Maria Osmala

Regularized modelling of dependencies between gene expression and metabolomics data in studying metabolic regulation

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 21.11.2011

Thesis supervisor:

Prof. Samuel Kaski

Thesis instructor:

D.Sc. (Tech.) Jarkko Salojärvi

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Maria Osmala

Title: Regularized modelling of dependencies between gene expression and metabolomics data in studying metabolic regulation

Date: 21.11.2011 Language: English Number of pages:10+156

Department of Information and Computer Science

Professorship: Information and Computer Science Code: T-61

Supervisor: Prof. Samuel Kaski

Instructor: D.Sc. (Tech.) Jarkko Salojärvi

Fusing different high-throughput data sources is an effective way to reveal functions of unknown genes, as well as regulatory relationships between biological components such as genes and metabolites. Dependencies between biological components functioning in the different layers of biological regulation can be investigated using canonical correlation analysis (CCA). However, the properties of the high-throughput bioinformatics data induce many challenges to data analysis: the sample size is often insufficient compared to the dimensionality of the data, and the data pose multicollinearity due to, for example, co-expressed and co-regulated genes. Therefore, a regularized version of classical CCA has been adopted. An alternative way of introducing regularization to statistical models is to perform Bayesian data analysis with suitable priors. In this thesis, the performance of a new variant of Bayesian CCA called gsCCA is compared to a classical ridge regression regularized CCA (rrCCA) in revealing relevant information shared between two high-throughput data sets. The gsCCA produces a partly similar regulatory effect as the classical CCA but, in addition, the gsCCA introduces a new type of regularization to the data covariance matrices. Both CCA methods are applied to gene expression and metabolic concentration measurements obtained from an oxidative-stress tolerant *Arabidopsis thaliana* ecotype Col-0, and an oxidative-stress sensitive mutant *rcd1* as time series under ozone exposure and in a control condition. The aim of this work is to reveal new regulatory mechanisms in the oxidative stress signalling in plants. For the both methods, rrCCA and gsCCA, the thesis illustrates their potential to reveal both already known and new regulatory mechanisms in *Arabidopsis thaliana* oxidative stress signalling.

Keywords: Bayesian modelling, bioinformatics, canonical correlation analysis (CCA), data fusion, dependency modelling, regularization

AALTO-YLIOPISTO
SÄHKTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Maria Osmala

Työn nimi: Regularisoitu riippuvuuksien mallintaminen geeniekpressio- ja metabolomiikkadatan välillä metabolian säätelyn tutkimuksessa

Päivämäärä: 21.11.2011 Kieli: Englanti Sivumäärä:10+156

Tietojenkäsittelytieteen laitos

Professuuri: Tietojenkäsittelytiede Koodi: T-61

Valvoja: Prof. Samuel Kaski

Ohjaaja: TkT Jarkko Salojärvi

Bioinformatiikassa eri tyyppisten mittausaineistojen yhdistäminen on tehokas tapa selvittää tuntemattomien geenien toiminnallisuutta sekä säätelyvuorovaikutuksia eri biologisten komponenttien, kuten geenien ja metaboliittien, välillä. Riippuvuuksia eri biologisilla säätelytasoilla toimivien komponenttien välillä voidaan tutkia kanonisella korrelaatioanalyysillä (canonical correlation analysis, CCA). Bioinformatiikan tietoaineistot aiheuttavat kuitenkin monia haasteita data-analyysille: näytteiden määrä on usein riittämätön verrattuna aineiston piirteiden määrään, ja aineisto on multikollineaarista johtuen esim. yhdessä säädellyistä ja ilmentyvistä geeneistä. Tästä syystä usein käytetään regularisoitua versiota kanonisesta korrelaatioanalyysistä aineiston tilastolliseen analysointiin. Vaihtoehto regularisoidulle analyysille on bayesiläinen lähestymistapa yhdessä sopivien prioorioletuksien kanssa. Tässä diplomityössä tutkitaan ja vertaillaan uuden bayesiläisen CCA:n sekä klassisen harjanneregressio-regularisoidun CCA:n kykyä löytää oleellinen jaettu informaatio kahden bioinformatiikka-tietoaineiston välillä. Uuden bayesiläisen menetelmän nimi on ryhmittäin harva kanoninen korrelaatioanalyysi. Ryhmittäin harva CCA tuottaa samanlaisen regularisointivaikutukseen kuin harjanneregressio-CCA, mutta lisäksi uusi menetelmä regularisoi tietoaineistojen kovarianssimatriiseja uudella tavalla. Molempia CCA-menetelmiä sovelletaan geenien ilmentymisaineistoon ja metaboliittien konsentraatioaineistoon, jotka on mitattu *Arabidopsis thaliana*:n hapetus-stressiä sietävästä ekotyypistä Col-0 ja hapetus-stressille herkästä *rcd1* mutantista aika-sarjana, sekä otsoni-altistuksessa että kontrolliolosuhteissa. Diplomityö havainnollistaa harjanneregressio-CCA:n ja ryhmittäin harvan CCA:n kykyä paljastaa jo tunnettuja ja mahdollisesti uusia säätelymekanismeja geenien ja metabolittien välillä kasvisolujen viestinnässä hapettavan stressin aikana.

Avainsanat: aineistojen välinen riippuvuus, bayesiläinen mallintaminen, bioinformatiikka, kanoninen korrelaatioanalyysi (CCA), regularisointi, tietolähteiden yhdistäminen

Preface

This thesis was done in Aalto University's Department of Information and Computer Science and at the Helsinki Institute of Information Technology (HIIT) in Statistical Machine Learning and Bioinformatics research group of prof. Samuel Kaski. The work was part of the Multibio project funded by Tekes and conducted in collaboration with the Helsinki University's Plant Stress research group led by prof. Jaakko Kangasjärvi. The work was mainly funded by Academy of Finland in the Adaptive Informatics Research Centre, an Academy Centre of Excellence on period 2006-2011, and partly by Tekes in the Multibio project.

I would like to thank professor Samuel Kaski and D. Sc. (Tech.) Jarkko Salojärvi for their instructions and supervision. I would also like to thank all plant biologists in the Kangasjärvi's Plant Stress Group, especially Tiina Blomster. I also thank Nina Sipari for a kind help and valuable discussions. Tiina and Nina measured all the data analyzed in this thesis.

Special thanks goes to Pejman Mohammadi, who initialized the work, Seppo Virtanen, the implementor of gsCCA, Seppo and Dr. Arto Klami for all valuable pieces of advice and discussions related to CCA, and Juuso Parkkinen for introduction to the eye diagrams. I also want to thank all my other collegues in the MI group for creating an inspiring working environment.

I would like to thank my dear friends and family just for their existence and support during my studies in TKK/Aalto. Jouni, thank you for always being there for me, especially during the ups and downs of the thesis making process. Finally, I want to thank the Father in Heaven for His mercy and answering prayers.

Otaniemi, November 21, 2011

Maria Osmala

Contents

Symbols, operators and abbreviations

Symbols

ρ_{xy}	Pearson's correlation
$\boldsymbol{\mu}$	mean of Gaussian distribution
σ^2	variance
$\boldsymbol{\Psi}_x$, $\boldsymbol{\Psi}_y$	sample covariance matrices
$\mathcal{N}$	Gaussian distribution
$\mathcal{IW}$	inverse-Wishart distribution
$\mathcal{IG}$	inverse-Gamma distribution
$\boldsymbol{a}_i$, $\boldsymbol{b}_i$	ith projection vectors in CCA
$\boldsymbol{A}$, $\boldsymbol{B}$	projection matrices in CCA
$\boldsymbol{B}_x$, $\boldsymbol{B}_y$	transformation matrices for the data set-specific latent variables
d	dimension
i	index over samples
j	index over variables
k	regularization parameter
$KL_d(p\|\|q)$	Kullback-Leibler divergence
$\mathcal{L}(q)$	The lower bound of the log marginal probability of likelihood
p	number of features in data set $\boldsymbol{X}$
$p(x)$	probability distribution
$p(x, y)$	joint distribution
$p(x\|y)$	conditional distribution
q	number of features in data set $\boldsymbol{Y}$
n, N	number of samples
r_{xy}	Pearson's sample correlation coefficient
R	canonical correlation
R^2	squared canonical correlation
s_x, s_y	sample standard deviation
$\boldsymbol{S}$	sample covariance matrix
$\boldsymbol{u}_i, \boldsymbol{v}_i$	ith canonical variate pair
$\boldsymbol{U}$, $\boldsymbol{V}$	canonical variates
$\boldsymbol{W}_x$, $\boldsymbol{W}_x$	CCA transformation matrices for the shared latent variables
x	sample, observation
X	random variable
$\boldsymbol{x}$	data vector
$\boldsymbol{X}$	data matrix
$\boldsymbol{z}$	shared latent variable
$\boldsymbol{z}_x$	latent variable specific to data set $\boldsymbol{X}$
$\boldsymbol{z}_y$	latent variable specific to data set $\boldsymbol{Y}$

Operators

$\| \bullet \|^2$	quadratic norm
$\int$	improper integral
$\hat{\bullet}$	maximum likelihood estimator
$\bullet^T$	matrix transpose
$\bullet^{-1}$	matrix inverse
$\mathrm{Cov}(\hat{\bullet})$	covariance of ML-estimate
$\mathbb{E}[L_1^2]$	mean squared error of ML-estimate
$\mathrm{Trace}(\boldsymbol{X})$	sum of the diagonal elements of matrix $\boldsymbol{X}$
$\hat{\bullet}^*$	regularized ML-estimate
$\overline{\bullet}$	mean
$corr(X, Y)$	Pearson's correlation
$I(X, Y)$	mutual information

Abbreviations

ABA	abscisic acid
ARD	automatic relevance determination
AUC	area under curve
BCCA	Bayesian CCA
CCA	canonical correlation analysis
Col-0	*Arabidopsis thaliana* ecotype Columbia-0, wild type
cDNA	complementary DNA
DNA	deoxyribonucleic acid
ET	ethylene
FDR	false discovery rate
GC	gas chromatography
GO	gene ontology
gsCCA	group sparsity CCA
HR	hypersensitivity response
i.i.d	independent and identically distributed
JA	jasmonate
KL	Kullback-Leibler
LC	liquid chromatography
MAP	maximum a posteriori
ML	maximum likelihood
mRNA	messenger ribonucleic acid
MS	mass spectrometry/spectrometer
PCD	programmed cell death, apoptosis
PR curve	precision-recall curve
P/R ratio	protein-per-mRNA ratio
RCD1	radical cell death induced protein 1
RCD1	radical cell death induced gene 1
rcd1	the mutant of gene *RCD1*
ROS	reactive oxygen species
SA	salicylic acid
rrCCA	ridge regression regularized CCA
TF	transcription factor

1 Introduction

Research in molecular biology has identified cellular components such as genes, proteins, enzymes and metabolites which function together in a highly regulated manner to form a complex system, life. In biological research, high-throughput technologies, such as DNA microarrays and mass spectrometric methods, are widely used to study the abundance and expression of these biological components in living organisms. High-throughput technologies are used to quantify the biological components in the whole genome or organism scale, leading to various *omics* data sets such as metabolomics and transcriptomics. The constantly increasing amount of various omics data types has motivated the development of sophisticated bioinformatics data analysis methods to dissect these data sets [1]. However, omics data sets obtained by high-throughput technologies pose several challenges to data analysis: the measurement technologies are highly inaccurate; differences in individual organisms cause variation in data; there is a difficulty in analyzing data with small sample size compared to the large number of features such as genes; and multicollinearity due to, for example, co-regulated and co-expressed genes makes the interpretation of the results hard.

There has been a growing interest in observing different types of high-throughput data sets jointly, and performing an integrated analysis on the data sets obtained from the same experimental units. The integrated analysis is justified because the regulation of the biological systems occurs on different biological components, and on different layers such as gene expression, translation, mRNA and protein degradation, enzyme activation, or the concentration levels of metabolites. Examples of this can be found in [2].

Systems biology as a research discipline focuses on interactions between the cellular components, as well as the functional mechanisms and behaviour of living systems arisen from these interactions [3–5]. The ultimate goal in systems biology is to understand the whole function of a cell in a systematic way, and to build *in silico* models to simulate life on computers. Computational systems biology attempts to analyse the regulation of a biological system as a whole by fusing different data sources, and by computational and statistical methods capable of handling the enormous data sets. In systems biology studies, the system under study is usually perturbed and high-throughput measurements are made to identify the components and their interactions. Moreover, integration of different data sets is often done given the biological domain knowledge that has been accumulated into the bioinformatics databases. An ever increasing amount of various biological high-throughput data sets is expected to emerge in the future, including single-nucleotide polymorphism (SNP) profiles, DNA copy number variation measured by array comparative hybridization methods [6], epigenetic modifications measured by chromatin immunoprecipitation combined with sequencing (ChIP-seq), and gene expression data measured using sequencing methods. Therefore, there is a growing need of computational systems biology and data integration methods.

The microarrays and mass spectroscopic methods used to quantify gene expression and metabolite concentrations, respectively, are probably the most widely used

high-throughput methods. They are the most suitable techniques for obtaining genome- or organism-scale measurements, and also the most useful in the analysis of metabolism and its regulation. The flow of information in metabolic regulation begins when the genes related to a certain metabolic pathway are transcribed to messenger RNA molecules (mRNA). The mRNA molecules are transferred to ribosomes in which the protein synthesis occurs according to the instructions contained in the mRNA molecules. This expression of mRNA molecules is a central functional and regulatory layer that determines the proteins produced in a cell [7]. Some of the proteins are enzymes catalysing biochemical reactions in which the metabolites participate. These reactions constitute a complex network of interactions between biomolecules which can be investigated by metabolomics techniques [2, 8]. Cellular metabolism can be seen as a result of the regulation occurring in the other layers of biological components such as transcripts. Therefore, understanding the metabolic network and its functional properties is crucial in deciphering the function of a cell [9,10]. The regulatory interconnection between gene expression and metabolism has led to the development of various methods which aim to integrate these two data sources [11–22].

This thesis concentrates on integrating high-throughput bioinformatics data using methods derived from canonical correlation analysis (CCA). Canonical correlation analysis [23] is a widely used method to fuse two or more paired data sets, i.e. distinct sets of features measured from the same samples. CCA seeks a linear representation of two or several paired data sets so that the representations are maximally correlated with each other, thus finding the statistical dependencies between the sets of features. Canonical correlation analysis is a symmetric method contrasting with asymmetric methods such as linear regression used to predict target value from input values. In CCA, both sets of variables can be considered as target and input variables simultaneously. Canonical correlation analysis fits to situations in which it can be assumed that the variation in one set of variables correlates with the variation in the other set and vice versa. Therefore, canonical correlation analysis suits well for analysing gene expression and metabolomics data because the changes in gene expression levels often regulate the metabolite concentration levels but, in addition, the metabolites might as well regulate the expression of some genes. Moreover, it is believed, in spite of the high dimensionality of the high-throughput data, that the interesting process underlying the data is in fact low dimensional. CCA performs a dimensionality reduction of high-dimensional data sets into lower dimensions for a more global view of the underlying biological process, and is thus a very attractive method for high-throughput bioinformatics data integration.

The noisy measurement techniques, the high dimensionality and small sample size, as well as the multicollinearity of the features typical to the high-throughput biological data pose various problems to data analysis. For example, canonical correlation analysis cannot be performed successfully when the number of features in both data sets exceeds the number of paired samples. This is because canonical correlation analysis requires the estimation of large covariance matrices and their inverses but the small sample size gives little statistical strength to obtain accurate estimates for the huge number of covariance parameters. As a result, the estimates

have very high variance, the covariance matrices are singular or ill-conditioned, and thus their inverses are unreliable, or cannot be computed at all. When the number of samples is low compared to the number of features, CCA model learned from the data overfits, i.e. starts to model irrelevant variation and noise in the data. Such challenges can be overcome by using classical regularization methods. For example, the classical ridge regression regularization has been adopted to canonical correlation analysis [24–27].

Applying the classical ridge regression regularization technique to canonical correlation analysis requires the optimization of the regularization parameters using, for example, cross-validation procedure. However, the optimization of the regularization parameters can be ambiguous and computationally very tedious. An alternative approach introducing regularization into statistical analysis is to apply Bayesian data analysis methods. Bayesian data analysis offers a number of advantages compared to the classical or frequentist regularization techniques such as addressing the uncertainty of the model parameters by assigning probability density to them; controlling the uncertainty in the modelling process itself; and the possibility to integrate over uninteresting model parameters. The Bayesian data analysis methods have an inherent regularization property that makes them particularly appropriate for complex bioinformatics data sets having small sample size and large number of features. The Bayesian data analysis methods have been applied in several publications to a single high-throughput data set (see references in [28]) but few exists for data fusion methods. The Bayesian framework can be applied to canonical correlation analysis by formalizing the model in a probabilistic framework [29–31]. In the probabilistic interpretation of CCA, the data is assumed to arise from a probabilistic generative model with a shared latent space for the two observed data sets. In the Bayesian framework, prior distributions are introduced to the model parameters. By selection of appropriate priors, the learned models become automatically regularized. For example, introducing Gaussian priors for the model parameters corresponds to the ridge regression regularization. Moreover, instead of using a part of the data to learn the model and another part to determine the model complexity, as is done in the cross-validation procedure to find the right regularization parameters for the classical regularized CCA, in the Bayesian framework, the whole data set is used to learn the model and its complexity jointly.

Bayesian canonical correlation analysis to fuse two high-throughput data sets has been introduced by Huopaniemi et al. [32]. The authors solve the problems of high dimensionality, small sample size and multicollinearity of the data by assuming that the data includes clusters of highly correlated and similarly behaving variables, thus decreasing the dimensionality of the variables. In contrast, this thesis considers a new variant of the Bayesian canonical correlation analysis [33] which models the individual genes and metabolites, not the clusters of them as in [32]. In this new Bayesian CCA, the data covariance matrices are replaced by low-rank approximations of them. This makes the rows and columns of a single covariance matrix dependent of each other, thus decreasing the number of effective parameters to be learned in the model. The low-rank approximation of the covariance matrices is obtained by introducing also source specific latent variables to the model, and by

applying an automatic relevance determination (ARD) prior [34] to the columns of the transformation matrix from the data set-specific latent space to the observed variable space. The data set-specific latent variables and the CCA components are learnt jointly while training the model. The model is called Bayesian CCA via group sparsity (gsCCA), and is inferred by variational approximation techniques [33].

No canonical correlation analysis method, which imposes this type of restrictions to the covariance matrix, has been applied to fusing gene expression and metabolomics data. The problem of small sample size and large number of variables has been solved either imposing regularization or sparsity to the CCA projection matrices [35–38] or assuming that the covariance matrix of the data is diagonal [35,38,39]. However, assuming diagonal matrix in canonical correlation analysis converts the method in practise to a principal component analysis (PCA) which cannot correctly model the shared variation between the two data sets.

This thesis compares the performance of the ridge regression regularized CCA (rrCCA) and the Bayesian CCA via group sparsity (gsCCA) in revealing relevant information shared between two high-throughput data sets. The performance is investigated by studying the generalization of the methods to unseen data by comparing the canonical correlations obtained by the two methods, and by assessing the variation in the data explained by different canonical components. The two methods are applied to data from a plant biology study in which a wild type *Arabidopsis thaliana*, Columbia-0, and an ozone-sensitive mutant *rcd1* were exposed to an oxidative stress. The gene expression and metabolomics data were measured from the same biological samples, and for both genotypes in a control condition and under ozone exposure at six time points starting from the exposure. The performance of the two CCA methods in finding interesting variation in the data is examined. Also a new gene-set enrichment analysis method is presented for more global analysis of the results. The study will hopefully reveal genes regulating the changes in metabolism (and vice versa), and suggest new hypotheses about the oxidative stress signalling in plants.

The thesis is structured as follows: Chapter 2 provides a necessary background of regulation mechanisms in biological processes, especially metabolism, and presents the biological research question relating to the oxidative stress signalling in plants. The oxidative stress signaling is presented, and the results of this thesis are analyzed in the light of this description. Chapter 3 introduces statistical and mathematical modelling concepts required in this thesis. The new Bayesian canonical correlation analysis with low-rank covariance matrix approximation for integrating gene expression and metabolomics data is given in Chapter 4. Chapter 4 also describes the methods to perform the comparison between regularized CCA and Bayesian CCA, and describes the biological data. Results for the data analysis are presented in Chapter 5. Finally, Chapter 6 concludes the thesis and discusses suggestions for future research.

2 Regulation of biological processes

Cellular functions result from tightly regulated biomolecular networks that control the cell. Biological regulation occurs at several layers of biological processes listed below, and some of them are shown in Figure 1:

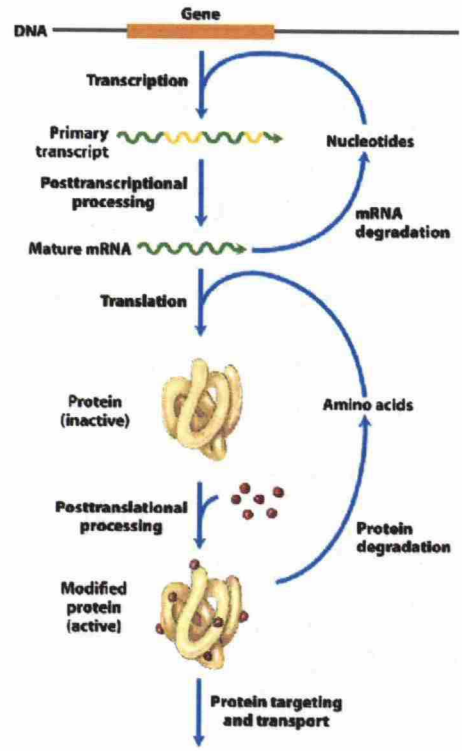


Figure 1: Layers of biological regulation. [40]

- synthesis of the primary messenger RNA (mRNA) transcripts, i.e. transcription
- post-transcriptional modification of the mRNA
- mRNA degradation
- protein synthesis i.e. translation
- post-translational modification of proteins to activate or inactivate them
- protein targeting and transport
- protein degradation
- metabolic regulation of enzymes

As can be seen in Figure 1, the genetic information to manufacture a protein is transferred from the nucleus to the cytosol by a transfer molecule called messenger ribonucleid acid (mRNA). The mRNA molecules are further translated to proteins that are the structural and functional units of the cell. This process is referred to as the central dogma in molecular biology [41]. Some of the proteins are enzymes that catalyse biochemical reactions in a cell. The enzymes regulate the activity of the individual biochemical reactions, and eventually the concentration levels of metabolites that are the substrates and end products of the reactions. Due to these hierarchical layers of biological regulation, cellular functions cannot be fully understood by studying one type of biological components at a time, for example, mRNA molecules, but through a comprehensive integration of the entire molecular machinery controlling the cell.

This chapter concentrates on biological regulation in eukaryotic organisms, especially at the level of transcription and metabolite concentrations. In addition, the technologies used to quantify the transcriptome and metabolome on a large scale are introduced. Transcriptomics and metabolomics reflect the regulation occurring at the two extreme levels of the information flow presented in Figure 1. However, much regulation occurs between the transcript and metabolite levels, including post-transcriptional and post-transcriptional regulation, as well as the regulation of enzymes, which have not been quantified in this work. Therefore, the details of biological regulation between transcripts and metabolites are not covered in this thesis. Despite this simplification, the methods discussed in this thesis are expected to reveal regulatory relationships between metabolites and genes. It is assumed that the expression level of a certain gene reflects directly the activity of the corresponding gene product which catalyses a metabolite reaction, and thus influences the levels of certain metabolites. Moreover, the changes in the metabolite concentrations can directly affect gene expression. Justification and critique of these assumptions are further discussed in the subsequent chapters.

2.1 Regulation of transcription

Genomics is a discipline in genetics concerning the study of genomes of organisms and the information contained therein. Genetics, systems biology and bioinformatics studies of the whole genome of an organism have given rise to various *genomics* data sets such as genome sequences, single-nucleotide polymorphism (SNP) profiles or DNA copy-number variation data.

The genome of an organism contains genes, the basic units of genetic information, which determine the instructions to manufacture proteins. A gene is defined as a locatable region of genomic sequence, corresponding to a unit of inheritance which is associated with regulatory regions, transcribed regions and/or other functional sequence regions [42]. A gene consists of two parts: a coding sequence, and regulatory regions. The coding sequence carries instructions for protein synthesis. Regulatory regions are sequences controlling the activity of a gene. A regulatory region proximal to a gene-coding region is called a promoter. Promoters are sites found near the transcription starting site (TSS) which is the sequence recognized

and bound by an enzyme catalysing the mRNA transcription, the RNA polymerase. The regulatory regions can also be distal, i.e. thousands of base pairs away from the promoter, and they can be situated both upstream or downstream of a promoter; they can reside even within the gene-coding region itself [40].

A gene is active or expressed in a cell when transcribed to a messenger RNA (mRNA). The complete collection of mRNA molecules of a cell or an organism is called a *transcriptome*. The main regulation of biological processes is believed to occur at the level of gene expression. From the perspective of energy efficiency, synthesizing mRNA molecules but not using them in protein synthesis would be a waste of resources. Usually the expression of multiple genes with interdependent activities is simultaneously regulated, requiring a synchronized control of transcription initiation. Transcripts that increase (decrease) in concentration under particular molecular circumstances are referred to as inducible (repressible). The process of increasing (decreasing) gene expression is referred to as induction (repression). For more details, see [40].

The activity or inactivity of the regulatory regions of a certain gene, and therefore the amount of expression of the gene, is determined by binding of regulatory proteins to the regulatory regions. The regulatory proteins include transcription factors (TF) of which many work by enhancing or interfering the interaction between the RNA polymerase and the promoter. Some of the transcription factors are repressors that impede the access of RNA polymerase to a certain promoter; this leads to negative regulation. On the other hand, a transcription factor can be activating, enhancing the RNA polymerase-promoter interaction, and hence the transcriptional rate. There are three types of transcription factors: sequence-specific factors, co-regulating factors, and factors that are part of the RNA polymerase machinery. Sequence-specific factors have DNA-binding domains that interact closely and specifically with the DNA. These binding domains usually include one or several recognizable and characteristic structural motifs. These structural motifs of TFs often recognize a certain sequence motif in the regulatory region. Sequence-specific factors function mainly by recruitment of transcriptional co-regulators to the regulatory region via protein-protein interactions that occur through specific interaction domains of the proteins. Co-regulators interact directly and indirectly with the components of RNA polymerase to regulate the activity of the RNA polymerase transcriptional machinery at the promoter. As a conclusion, the DNA binding transcription factors interpret the genetic regulatory information, and transmit the appropriate response through co-regulators to the RNA polymerase transcriptional machinery. More details can be found in [40, 43].

The eukaryotic promoters are generally inactive in the absence of regulatory proteins because the storage of DNA within the chromosomes effectively renders most promoters inaccessible. Moreover, the RNA polymerase usually requires an array of transcription factors in order to bind to a promoter and start the transcription. At the promoter, there is usually a cluster of sequence motifs, i.e. the recognition sites for multiple sequence-specific TFs. Multiple transcription factors can bind to a single regulatory region, and form a protein complex, and their co-operation is what finally determines the regulatory state. The binding of several regulatory pro-

teins improves the specificity for transcriptional regulation. Specificity is thought to be required by large eukaryotic genomes, as in large genomes a single specific binding sequence of a certain transcription factor will occur randomly with a large probability [40].

Increase in the concentration levels of transcription factors in the nucleus, or activation of transcription factors by intracellular or extracellular signals, triggers their regulatory effects. The increase in the concentration level of a TF in the nucleus might result from the induction of the transcription factor-coding gene, or from the transport of transcription factors from the cytosol to the nucleus. The former can be observed as a higher expression of TF genes. A transcription factor might, however, be inactive (active) just after translation but might be later activated (inactivated) by post-translational modifications such as phosphorylation, or binding of a signal molecule. TF transportation or post-translational activation are common end results of signal transduction pathways, which transfer signals from intracellular and extracellular parts of the cell to the nucleus where they affect the transcription as a respond to the signal (examples of this can be found in [40]). This change in the activity of transcription factors cannot be seen in the changes of expression of the corresponding genes.

Binding of specific regulatory proteins to the regulatory regions of a gene might switch it to a functional and active form leading to the transcription of the gene. Transcription starts when the double-stranded DNA is opened in the proximity of the gene-coding sequence, and the RNA polymerase binds to the promoter. The polymerase enzyme reads the protein-coding sequence of the gene, and produces a complementary pre-mRNA molecule. After post-transcriptional processing of the pre-mRNA, the mature mRNA molecule is exported from the nucleus to the cytosol. The complete collection of transcribed mRNA molecules in genome scale, the transcriptome, is a central functional layer of the genome regulating protein production in a cell, and therefore also the major part of the biological processes [44]. Different collections of genes are active in different contexts: transcriptional activity varies by cell type, environmental conditions and time. It is believed that at a given time most genes are expressed at low concentration, perhaps only one or few copies of the mRNA per cell on average, whereas a small number of genes are highly expressed, potentially with thousands of copies per cell [45, 46]. The mRNA molecules are not everlasting: they are degraded back to nucleotides when not needed anymore. A particular transcript might not ever be translated to a protein if it is chemically unstable or condemned to degradation. It has been shown that the degradation rates of transcripts are highly regulated, and affect the number of protein translated from mRNA molecules [47–49].

2.1.1 Quantification of the transcriptome

Transcriptional profiling is currently the main high-throughput technique used to investigate gene functions at genome- and organism-wide scale [50, 51]. The transcriptome is measured by DNA microarrays [52–54] which allow to routinely measure the expression levels of tens of thousands of mRNA transcripts in a given biological

sample [55]; the number of mRNA abundances to be analyzed may be as large as 40,000.

The genomic DNA in chromosomes consists of two long polymers of simple units called nucleotides. There are four types of nucleotides in the DNA: thymine (T), adenine (A), cytosine (C) and guanine (G). The two DNA strands form a double helix structure in which each nucleotide on one strand interacts with just one nucleotide on the other strand. This is called complementary base pairing. Nucleotide A forms base pairs with T and C form base pairs with G. The two strands forming the double helix are called complementary to each other. The base pairing property of nucleotides is behind the DNA microarray technology; a microarray slide contains complementary sequences to the mRNA molecules residing in a biological sample. In the microarray analysis workflow, first mRNA from a biological sample, or complementary DNA molecules obtained from the mRNA using reverse transcriptase enzyme, are marked with fluorescence labels. Then the sample is injected on a microarray slide, and the mRNA molecules contained in the sample are bound to the complementary nucleotide sequences on the slide, i.e. spots or probes. This is called hybridization. Each probe on a microarray slide is expected to uniquely hybridize with its intended target sequence in the sample. The activity of a certain transcript can be detected by investigating its hybridization level, reflecting the target mRNA concentration in the sample. The hybridization level is estimated by measuring the intensity of light emitted by the fluorescent labelled mRNA molecules with a laser scanner. For more details, see [52, 54].

The DNA microarray slides can be of three different types: complementary DNA (cDNA) arrays, oligonucleotide arrays [52,54,56,57] or bead arrays. Oligonucleotide microarrays are most widely used, and they are shown to be more accurate than cDNA arrays [52,54]. The data analyzed in this thesis was measured using custom-made cDNA microarrays, so the manufacturing technology, experimental procedure, data pre-processing and analysis is covered in this thesis only for cDNA arrays. The data analyzed in this work was obtained as already pre-processed and normalized, so the data pre-processing is not covered in detail.

The cDNA microarrays are manufactured by attaching oligonucleotides or cDNA molecules on a class slide by a robotic arrayer. The robotic arrayer uses multiple pins to pick up solution of cDNA molecules, and then deposites them on a glass slide. Therefore, cDNA arrays are called contact-spotted microarrays. Other techniques to produce cDNA arrays are ink-jet deposition [58], and electrophoretically driven deposition [59]. The probes can be synthesised, or cDNA can be generated via reverse-transcription of all mRNAs from an organism under study to obtain so-called expressed sequence tags (ESTs) which are then amplified by polymerase chain reaction (PCR), purified, and spotted on the microarray slides. The technique is versatile, and probes of several hundred to thousand base pairs in length can be attached to the slide. The cDNA platforms are usually two-colour microarrays, indicating that RNA or cDNA obtained from differently treated samples are labelled with different fluorescent cyanide (Cy) markers: Cy5 being fluorescent in red region, and Cy3 being fluorescent in green region. Two-colour microarrays are always used to compare the expression ratio between differently treated samples of which the

other is usually the control sample. This direct comparison reduces the noise level in the final data. The cDNA microarrays tend to have a large variation in spot size, and morphology between spots and between arrays, thus making it more difficult to compare two samples hybridized on different arrays. Comparing samples directly on the array decreases the noise due to the inefficient spot detection. Furthermore, the same sample labelled with different fluorescent markers may show different hybridization affinity to the slide which might also vary between genes. To reduce this dye effect, a dye-swap is commonly performed. In dye swap, the colouring order of the samples is swapped to address the uncertainty associated with the binding affinity of differently labelled samples. For more details, see [52, 54].

After the samples have been hybridized onto the microarrays, a laser scanner detects the abundance of bound mRNA or cDNA molecules. For two-colour microarrays, the abundance of both colours is measured separately resulting in different channels. The quantification assigns signal intensity values to each spot or probe. The relative gene expression values obtained from cDNA experiment are commonly transformed to a logarithmic scale with base 2. The principal motivation for this transformation is to make variation roughly comparable among measures that span several orders of magnitude.

The next step in the microarray data analysis is usually normalization to remove irrelevant variation, and preserve the interesting variation. Interesting variation is usually the biological variation. Differences of individuals in biological samples, and inability to control the biological experiment also lead to irrelevant biological variation. Differences in the scale intensity between red and green channels, different labelling efficacy of the samples, different hybridization times and conditions, as well as different scanning sensitivity and laser power all lead to unwanted technical variation [60]. The normalization method to remove irrelevant biological and technical variation used for the data analyzed in this thesis is presented in Chapter 4.2.2.

Despite the huge advances in microarray technology and their popularity in biological research, caution is needed when applying them. In general, the microarray technology is still to a large extent irreproducible between microarray experiments done at different sites, at different platforms and between laboratories. The irreproducibility aspect of microarray technology is addressed in the editorial [61] preceding the report summarizing the large-scale Microarray Quality Control Project [62]. After 15 years of research and development, broad consensus is still lacking concerning best practice not only for experimental design and sample preparation but also for data acquisition, statistical analysis and interpretation. The microarrays measure the abundance of the mRNA molecules in a cell at a certain time instance but the concentrations of the mRNA molecules do not tell much about the transcriptional activity of genes, i.e. the rate at which a gene is transcribed to a messenger RNA [63]. This is because microarray technology does not address the post-transcriptional mRNA stability, activity and degradation rate which are tightly regulated, and vary between genes and due to differences in cell environment. [47–49,64,65]. Moreover, it has also been observed that the mRNA half-lives of some genes cluster into comparatively tight groups [63]. This has given rise to a concern that the co-expression of genes observed in the results of microarray experiments might not reflect the tran-

scriptional co-regulation, as is traditionally interpreted, but would instead result from the clustering of mRNA degradation rates [63].

2.1.2 Statistical significance of differentially expressed genes

After obtaining the normalized intensity values of the spots or probes, the significance of differentially expressed genes, for example, between control samples and treated samples, is determined. The genes can be differentially expressed in different subjects exposed to varying physical and biochemical conditions. Identification of genes that display tendencies of increased or reduced expression helps to discover the functional roles of genes in an organism, to group genes according to common functions, or to understand the relationships among genes in a biological system. Usually only genes having intensity values or fold changes in a log 2 scale compared to a control sample above 1 or below -1 are investigated. A standard t-test is usually performed for each gene to find out genes having expression significantly different due to a treatment compared to a control condition.

The statistical test gives p values for each gene, which can be investigated for statistical significance. However, obtaining p values for tens of thousands of genes results in a problem; for 10000 significantly expressed genes with a significance level of 0.05 leads to 500 false positive discoveries. This is called a multiple comparisons problem [66]. The large number of false positive results obtained in a multiple comparison test can be avoided by setting the target significance threshold much lower than it would be for a single test. Another way is to use a multiple-hypothesis testing error measure called a false discovery rate [67]. The false discovery rate (FDR) is the expected proportion of false positive findings among all the rejected null hypotheses. FDR analogue of the p value is called q value. The q value of an individual hypothesis test is the minimum FDR at which the test may be called significant. The q values can be directly estimated rather than fixing a level at which to control the FDR [68]. The calculation of q values for genes to determine their statistically significant expression is today a standard procedure in microarray data analysis.

2.1.3 Gene-set enrichment analysis

The microarray data analysis can be performed at three levels: at single gene level, at the level of multiple genes, or at the level of networks of genes [69]. The analysis of differential expression of a group of genes is more reasonable than the analysis of expression levels of single genes, because a set of multiple differentially expressed genes gives more statistical strength to the analysis of differences between biological samples than single genes. In addition, the analysis of sets of genes is well advised because there might be groups of genes that are co-regulated and therefore co-expressed, or otherwise related, that are induced or repressed together due to some stimuli. The list of differentially co-expressed genes can be further translated into a better understanding of the underlying biological phenomena. In particular, the results can be investigated in the context of gene or metabolic regulatory networks, or signalling pathways on the whole organism level. Methods for analyzing differentially

expressed gene sets are, for example, functional class scoring [70, 71], Fisher's exact test and gene-set enrichment analysis (GSEA) [72, 73]. In GSEA, the expression values of all transcripts all investigated, and the enrichment score for a gene set is calculated based on the order of the expression values. The statistical significance of the score is determined by a permutation test. GSEA finds significantly enriched gene sets among genes that are related to each other, although the individual genes might lack significantly differential expression.

Usage of gene ontology (GO) classes as predefined gene sets is a widely adopted method to study groups of differentially expressed genes [74, 75]. However, there are a number of limitations related to the type, quality and structure of the GO annotations available. First, all types of gene-set enrichment analysis have been criticized of being limited by the fact that each functional GO category is analyzed independently without a unifying analysis at a biological pathway or organism level [73, 76]. Second, the GO class enrichment methods do not suit for a systems biology approach that aims to account for system level dependencies and interactions between different biological components. An alternative way is to investigate the enrichment of biological pathways obtained from databases such as KEGG and BioCYC. Furthermore, there is no enrichment analysis which would take both differentially expressed genes and metabolites having differential concentration levels as an input, and give enriched pathways as a result.

Gene-set enrichment analysis gives small p values for enriched gene sets but if very many gene-set enrichment tests are performed, the p value correction, for example, using Benjamini-Hochberg method [67], is again required. In gene-set enrichment analysis, significantly up-regulated and down-regulated genes can be studied separately, or together by considering only the absolute fold changes of the genes. Moreover, usually only those genes belonging to at least one GO class are used to calculate the test statistics; this increases the p values, and leads to a smaller number of enriched gene sets.

Co-expressed groups of genes, if they in addition have similar functions, are likely to be regulated via the same regulatory mechanisms. They are most likely to have their promoter regions bound by a common transcription factor and share common regulatory motifs [77]. Allocco et al. [77] have investigated data from a genome wide binding analysis of transcription factors in combination with mRNA expression levels and existing functional annotations to quantify the likelihood that co-expressed genes, or genes sharing the same biological function, will be bound by a common transcription factor. The authors found that this effect is present at relatively high levels of expression similarity. Furthermore, they estimated the probability of two genes being bound by the common transcription factor as a function of correlation between the expression values of the genes. Especially, in order for two genes to have a greater than 50 % change of sharing a common transcription factor binding site, the correlation between their expression must be greater than 84 %. Seeking common potential transcription factor binding motifs from co-expressed genes, such as genes belonging to a same GO class, is one option for further analysis of the results obtained by gene-set enrichment analysis.

2.1.4 A new gene-set enrichment analysis based on area under precision-recall curve

A good gene-set enrichment analysis method should work even when the number of samples is small, and as a result it should give enriched gene sets containing a large number of genes. This thesis introduces a new type of gene-set enrichment analysis method that uses area under precision-recall curve to determine the gene-set enrichment score. Precision and recall are common concepts in information retrieval for evaluating retrieval (classification) performance [78,79]. In this new method, the absolute expression of all genes belonging to at least one gene-set annotation are listed in decreasing order, and precision and recall of each enriched gene ontology class are calculated on each gene in this list. While the genes in the list are gone through, the precision and recall for each class are calculated as follows:

$$\begin{aligned} \text{precision} &= \frac{\text{Genes belonging to a class encountered thus far}}{\text{The total number of genes encountered thus far}} \\ \text{recall} &= \frac{\text{Genes belonging to a class encountered thus far}}{\text{The number of genes belonging to the class}} \end{aligned} \tag{1}$$

The precision can be plotted as a function of the recall, and the area under the resulting curve (AUC) can be calculated. This area corresponds to the average precision of the gene-set enrichment test [80]. The obtained value could be compared to the distribution of AUC obtained by randomizing the order of the genes. A gene set could be considered significant if the AUC for that gene set is larger than 95 % over the AUCs obtained by randomization. This gene-set enrichment analysis method has the property of giving high significance for gene sets including large number of differentially expressed genes.

Precision-recall curves have the disadvantage of having a distinctive saw-tooth or "wiggled" shape. If the (k+1)*th* gene in a list do not belong to the GO class under study, its recall is the same as for k*th* gene but the precision drops. On the other hand, if the gene belongs to the class, then both precision and recall increase and the curve continues up and right. Solution to this is to interpolate the curve. Interpolated precision $p_{interp}(r)$ at recall r is the maximum precision over all recalls greater than r. The interpolated precision-recall curve is a step-curve, and area under it can be calculated easily. For more details, see [80–83]

2.1.5 Mutation and knock-out of a gene in microarray studies in revealing the function of the gene

Genetic studies usually rely on the assumption that disrupting the activity of a gene required for a process will have a specific effect on that process. Disrupting a gene by mutating it or by a knock-out is often used to reveal the unknown function of the gene. Correlations between perturbation in expression of one gene and changes in the activity of other genes imply that a functional interaction exists between them [84,85]. In a mutant organism, the concentrations of the directly affected gene products can be reduced or increased by variable degrees. The resulting changes in target gene expression are then interpreted in terms of activating or repressive

regulatory interactions between mutated and regulated genes. Up-regulation of a target gene in a mutant indicates that the mutated gene represses its target; down-regulation indicates activation. However, interpretation of mutant expression patterns is not always straightforward especially if the system is spatially distributed, and/or involves feedback regulation between more than two or three genes [86]. A specific limitation in the genetic studies is that it is often unclear from any evidence whether an interaction is direct or mediated via one or more intermediate regulators. The solution in such a situation is to mutate several genes.

2.2 Regulation of translation and protein levels

The mRNA molecules floating in the cytoplasm are used as instructions to manufacture proteins, the structural and functional units of the cell. The instructions are coded in segments of three nucleotides in the gene-coding region or the mRNA transcripts. These segments are called codons. The order of the codons in the gene-coding region or in the mRNA molecule determines the properties of the gene product.

A protein molecule is composed of amino acids that are arranged in a chain. The amino acid chain or sequence is called the primary structure of a protein, and the order of the amino acids in the chain is determined by the codon sequence in the mRNA molecule. Different codons correspond to one of 20 proteinogenic amino acids. There are also other amino acids which do not end up into protein sequences. The manufacture of the proteins, i.e. the protein synthesis or translation, occurs in special molecular machinery called a ribosome. In ribosome, the different amino acids that will be attached to each other are brought to the translating site by transfer-RNA molecules. When the amino acid sequence, i.e. the primary structure of a protein is completed, it often folds around itself to form secondary and tertiary structures. Moreover, different amino acid sequences can join to form multimer proteins; this is called a protein tertiary structure. A folded protein or enzyme may not be active immediately after translation; several molecules or ligands, such as sugars, metals, cofactors and vitamins, can bind to the protein to enable its activity and functionality, or the enzyme is inactive until a certain signalling molecule binds to it.

In general, translational regulation is achieved by regulating mRNA stability, translation initiation, elongation and termination [87, 88], and the overall rate of protein synthesis [89, 90]. The concentration level of a protein is determined by these mechanisms together with protein degradation. The protein degradation is highly specific and tightly regulated [91–94]. The abundance of protein molecules in a biological sample, i.e. *proteomics* data, can be measured by 2D gels, and by a western plot [95] together with mass spectrometric methods [96–98], or libraries of green fluorescent protein -tagged proteins can be determined [95, 99–102]. However, the proteomics data is difficult to measure in a high-throughput manner; the best approach thus far for large-scale quantification of protein levels is shotgun proteomics [103, 104].

Due to the difficulty of measuring the protein levels in a high-throughput man-

ner, the expression levels of transcripts are assumed to reflect the concentration of proteins and enzymes, and therefore the activity of the corresponding biological processes catalysed by them. Therefore, high correlation between transcriptome and proteome is desirable [105]. However, the mRNA concentration can only partially explain variation in protein concentration levels. The different processes occurring between transcription and appearance of proteins such as splicing; transportation of the mRNA molecule from the nucleus to the cytoplasm; mRNA maturation and editing; RNA interference; mRNA-binding proteins; the degradation rate of the mRNA molecules; the regulation of translation initiation, elongation, termination; and protein degradation raise a question whether the transcription levels reflect the concentrations of the final gene products, the proteins. Furthermore, noise in the measurement technologies, biological noise originating from the inherent stochasticity of biochemical processes [101, 106], difference across individual cells in a population, and subtle environmental differences and genetic mutations [107] cause inaccuracy in predicting protein levels from the gene expression measurements.

In yeast and bacteria, and to a lesser extent in animals and plants, there is a substantial and significant correlation between protein and mRNA concentrations. Typically 30-85 % of the variation in protein levels can be attributed to variation in mRNA expression [108]. The rest can be explained by post-transcriptional and post-translational regulation and by measurement errors. A recent review addressing this issue [109] summarises the current state of knowledge about large-scale measurements of absolute protein and mRNA expression levels, and the degree of correlation between these two measurements. Authors focus particularly to the protein-per-mRNA (P/R) ratio. If there were no regulation in the translation or mRNA degradation, P/R ratio would be identical for all genes. In reality this is not observed; P/R ratios vary widely for genes measured from one cellular sample to another. Moreover, P/R is different for different genes, and might change for a given gene under different conditions. Protein stability is one of the major factors determining P/R, and P/R often depends on the type of protein. For example, in fission yeast, the correlation between gene expression and protein levels is strong for kinases, cell cycle genes, signalling and metabolic proteins but weak for proteins which form complexes [110].

Several large-scale studies exist in which the changes in protein concentrations are compared to the changes in mRNA concentrations [111–114]. Correlation coefficients between mRNA and protein levels vary widely across organisms and are often surprisingly low. In bacteria, the squared Pearson's correlation coefficient between transcriptomics and proteomics ranges from 0.2 to 0.47, in yeast from 0.34 to 0.87 and in multi-cellular organisms from 0.09 to 0.46. A recent study of yeast compared not only the total mRNA levels but also the translationally active transcrips, i.e. those that are bound to ribosomes, to protein concentration measurements [115]. The most significant but not the highest squared correlation between protein and mRNA concentrations in yeast was 0.42. All in all, squared correlation for several recent measurements lies around 0.4. Therefore, 40 % of the variance can be explained by changes at the transcript level, and 60 % by other means. Multicellular organisms display on average the lowest correlation between protein and mRNA

levels. In multicellular organisms, the correlation is particularly poor for genes of signal transduction and transcriptional regulation, possibly due to extensive post-transcriptional control, or due to the low and error-prone concentrations of proteins and transcription levels. However, a study in *Arabidopsis thaliana* describes a good correlation between protein and mRNA levels, the Pearson's correlation coefficient has been observed to vary between 0.52 and 0.68, similar to the correlations observed for *C. elegans* and *D. melanogaster* [116]. Figure about the correlations between transcripts and protein levels in various organisms are shown in subfigures A., B. and C. of Figure 2. This figure describes also the observed correlation between transcripts and protein levels in different studies conducted over years for several organisms (subfigure D). The subfigure D in Figure 2 shows that the overall correlation between transcripts and proteomics in living world has in recent measurements been around 0.4 [109].

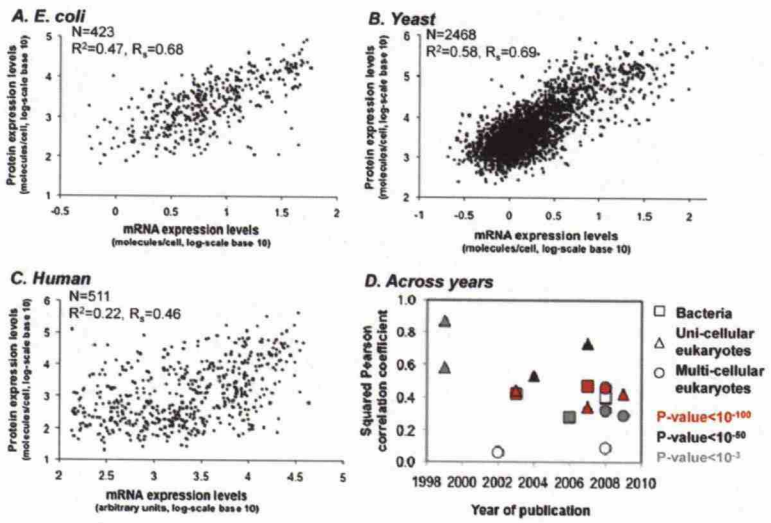


Figure 2: Correlation between transcriptome and proteome in various organisms. [109]

2.3 Regulation of enzyme activity

Some proteins are enzymes catalysing biochemical reactions in a cell. In biochemical reactions, small chemical substances, i.e. metabolites, are converted from one to other. Fast enzymatic transformations of metabolites cause a continuous flux of matter, and hence the dynamic properties of the metabolism. The rate at which the matter flows in a metabolic reaction, the flux, reflects the activity of an enzyme catalysing the reaction. The regulation of enzyme activities in metabolic pathways is by far the most common form of metabolic regulation. A reaction flux may simply increase by increased production of the corresponding enzyme resulting from the induction of the enzyme-coding gene. However, this is not the only mechanism to control the flux. First, the activity of a newly synthesised enzyme can be regulated by various processes: an enzyme may be bound by a certain molecule, making it

active, or activation may require that a piece of enzyme is split off from it. Second, phosphorylation is a common mechanism either activating or inactivating an enzyme. Third, an enzyme can be informed of the state of the metabolic pathways by specific chemical signals such as the concentration of substrate, product and specific regulators; binding of a certain molecule to an enzyme can change the electrical and structural properties of an enzymes leading, for example, to altered affinity to the substrate molecule. Fourth, many enzymes catalysing the first steps in a metabolic pathway are prone to inhibitory regulation of the end products of the pathway. An inhibitor is any molecule that reduces the velocity of an enzymatic reaction. If an enzyme of a metabolic pathway is inhibited by the pathway end product, the mechanism is called negative feedback inhibition.

The activities of enzymes are not directly available in this thesis. However, it is assumed that the gene expression levels reflect the activities of the proteins and enzymes the in metabolic pathways. Furthermore, enzymatic activities can be partly accessed through the metabolite concentrations because the metabolite concentrations are ultimately the product of the complex biological regulation. An increase or a decrease in metabolite levels may reflect activation or inactivation of the pathways producing or consuming them. Moreover, if a certain enzyme is inhibited by a certain metabolite, this metabolic down-regulation may also lead to transcriptional down-regulation because synthesizing an enzyme that will anyway be inhibited is waste of resources.

2.4 Regulation of metabolism

2.4.1 Metabolome

Metabolites are small molecules participating in the biochemical reactions, and are required for the maintenance, growth and the normal function of a cell. The complete set of metabolites in an organism is denoted as a *metabolome* which can be partly identified and quantified using various *metabolomics* technologies. The amount of metabolites in a living organism can be tens of thousands. For example, in plant kingdom where the estimated number of species is more than 400 000, the total number of metabolites is estimated to be between 2 000 000 – 1 000 000 [117]. A single *Arabidopsis thaliana* plant is expected to produce 5000 metabolites or even more. A database of genetic and molecular biology data for *Arabidopsis thaliana*, AraCyc (version 6) already contains 2,632 compound entries. This number encompasses the documented *Arabidopsis* metabolites, as well as those residing in manually curated metabolic pathways [117–119]. The huge number of metabolites in plants originates from the thousands of secondary metabolites. Secondary metabolites are molecules not essential to survival, including flower pigments, phenylpropanoids, flavonoids, terpenoids, glucosinolates and alkaloids. These have important function, for example, in cellular signalling and adaptation to abiotic and biotic stresses [120]. The complete set of metabolites is different in different parts of a plant because there are large differences in functions between tissues and the cells comprising them. This brings spatial heterogeneity to the metabolism. On the other hand, the same reac-

tions can occur in different compartments of a plant such as plastids, mitochondria, peroxisomes and vacuolar compartments. Indeed, many basic pathways like tricarboxylic acid cycle (TCA), glycolysis, oxidative pentose phosphate pathway (OPPP) and organic acid metabolism are present in more than one compartment [121].

2.4.2 Metabolic reaction activities define the phenotypic state of an organism

Metabolic reaction rates or, in other words, the flux distribution, are the manifestation of the regulation at enzyme and metabolite level. The rates of individual reactions in biochemical pathways ultimately define the phenotype of an organism. The production and consumption rates of the metabolites are directly connected to the reactions in metabolic pathways, and hence to the phenotype. The metabolites are more directly connected to the phenotypic state of a cell than, for example, transcriptomics. Moreover, the metabolism represents the amplification of regulation occurring in other biological layers such as transcripts and enzyme activities. Identifying and quantifying the metabolites under a specific condition, such as stress or gene knock-out, and comparing the levels to measurements obtained in a control condition can provide useful information of the regulatory processes in an organism. Indeed, metabolomics data has been used to determine the function of unknown genes [118, 122–127]. For example, by measuring the change in relative concentrations of metabolites as the result of the deletion or overexpression of a gene allows to locate a novel gene product on the metabolic map [128].

The guilty-by-association principle of gene expression [129] states that a set of genes involved in a certain process is generally co-regulated and thus co-expressed under the control of a shared regulatory system [118, 125, 126]. Therefore, if an unknown gene is co-expressed with known genes of a particular biological process, it is assumed that this unknown gene may be involved in this process. This co-occurrence principle can be extended to co-regulated metabolites as well, or expression pattern of genes associated with a particular metabolic pathway can be combined with co-accumulated metabolites also involved in the pathway. In this manner, regulatory relationships between genes and metabolites might be found.

2.4.3 The nature of metabolism

The biological functions of metabolism are, for example, to support growth, to synthesise and to turnover compounds, and to accumulate metabolites that have a role in coping with various stressful situations. The nature of metabolome distinguishes it from the other omes such as transcriptome and proteome. Metabolism in living organisms operates as a highly integrated framework [130–132]. Metabolites are not synthesised in isolation; rather, large sets of metabolites are often synthesised simultaneously and, therefore, the synthesis of one metabolite typically requires co-operation of many biosynthetical pathways. Metabolic pathways are also very branched. In conclusion, when the synthesis of a certain set of metabolites is up or down-regulated, it is often necessary to alter the fluxes and/or metabolite levels in several pathways. For more details, see [133].

Metabolism is highly redundant and provides several alternative routes in synthesizing a certain metabolite. Therefore, the decrease in the concentration level of one enzyme protein due to a knock-out of the corresponding gene or an environmental factor do not necessarily result on phenotypic differences in metabolome [134–138]. The biological reason for this might be to ensure stability of the metabolic network with respect to genetic mutations. Moreover, the reaction rates of pathways might not change in spite of the changes in the concentration levels of metabolites, or even changes in the activity of an enzyme catalysing the reaction. For example, the response to a decrease in the activity of an enzyme might be an increase in the concentration of substrates of that enzyme which ensures that the flux alters only slightly if not at all [139]. As a conclusion, metabolic networks are very robust to different kinds of perturbations. [140]. On the other hand, metabolome data might enable the detection of the phenotype of silent mutations. A silent mutation is a mutation in a genome which leads to no overt phenotype but which can be mapped by omics techniques such as metabolomics [141, 142].

For certain metabolic pathways, individual reactions are organized such that the metabolic intermediates constituting a pathway are always bound to the enzyme surface, and the products are handed on as substrates for the subsequent reactions. In such cases, intermediates of these reactions typically participate in one specific reaction, and are not detected in biological samples. On the other hand, central carbon pathways, such as glycolysis, gluconeogenesis, TCA and OPPP, are not organized in the same way. The intermediates of these pathways are indeed detectable. Moreover, the highly connected nature of metabolism requires that a number of compounds function as intermediates and precursors for more than one pathway. Such metabolites cannot be bound to enzymes because they always have to be available for more than one pathways, and more importantly, the rates of these pathways are likely to vary under different conditions [139]. These metabolites can be observed. Because the metabolites are constantly produced and consumed, the quantity that is observed by metabolomics techniques for intermediate metabolites in pathways is the difference between produced metabolites and consumed metabolites in a steady-state, i.e. when the reaction rates of metabolic pathways are assumed to be constant [120].

2.4.4 Metabolic regulation

The changes in the metabolite concentrations, i.e. the accumulation and depletion of metabolites, have several origins. For example, the accumulation of a certain metabolite can be due to its increased biosynthesis, or the down-regulation of the pathways having the accumulated metabolite as a starting compound. The accumulation of the metabolites functioning as precursors of several biomass components may also result from impaired growth [120]. On the other hand, depletion of a metabolite might occur due to the down-regulation of its biosynthesis or increased demand of that metabolite. However, the increased demand often intensifies also the synthesis of the metabolite, for example, by the negative feedback regulation mentioned in the previous chapter [143]. Accumulation and depletion can also be

affected by the degradation rate of the metabolite. Changes in the enzyme or pathway activities are not always visible in the metabolite concentration levels due to the coupling of the metabolic network, but for the same reason, even small perturbations in the proteome or enzyme activity can cause significant changes in metabolite concentrations. For more information, see [120].

The metabolic regulation covers the molecular mechanisms by which the activity of the individual enzymes or fluxes are modulated and controlled. Up- and down-regulation of enzymatic activities can be qualitatively estimated from the expression levels of enzyme genes and proteins, as well as using enzyme assays. In certain cases the pathway flux is controlled by the total amount of an enzyme present that is low enough to qualify as a rate-controlling step. In these cases, the regulation is visible in transcriptomics and/or proteomics data. In addition to regulation provided by transcriptomics and proteomics, metabolic regulation is another level of regulation of biological processes [144]. Metabolite levels themselves regulate various processes, for example, by negative feedback regulation. Therefore, certain compounds function both as metabolic intermediates and as metabolic regulators. The metabolic regulation can be investigated to some extent by the identification and quantification of metabolite levels.

The enzymatic regulation of metabolic pathways can occur in two different ways. First, the metabolic fluxes can be controlled by a small number of "key" regulatory or "rate-limiting" enzymes which function as bottlenecks in the biochemical pathways [145]. Second, the whole set of enzymes regulating several steps along a metabolic pathways can be regulated [146–148]. The bottleneck enzymes are usually regulating the first steps in biological pathways, and they are often controlled by post-transcriptional and metabolic regulation. The reaction catalysed by these enzymes is usually irreversible indicating that the reaction proceeds only in one certain direction [143, 149, 150]. The bottleneck enzymes are common, for example, in amino-acid biosynthesis pathways [139]. The activity of irreversible bottleneck enzymes regulated, for example, by post-transcriptional modifications, negative feedback loops or other metabolic regulation cannot be accessed directly by the data acquirement methods covered in this thesis.

The regulation of the whole set of enzymes along a metabolic pathway has been proposed to be the most effective way to alter flux. These enzymes often catalyse reversible reactions that can easily proceed forward and backward. Nevertheless, they can have a large effect on flux [149–153]. The activity of these enzymes may result from coordinated changes in gene expression, parallel post-transcriptional modifications of many enzymes, or recruitment of enzymes into complexes [131, 154]. Therefore, the regulation of these enzymes could be observed in the expression of corresponding genes. Moreover, if several enzymes along a metabolic pathway or belonging to several related pathways show similar response to perturbation in gene expression, it would give evidence of the co-regulation of these enzymes already in the transcriptional level.

The co-regulation of genes encoding the enzymes synthesizing reactions along a metabolic pathway may result from the binding of some transcription factors to the promoter regions of the genes. This type of a regulation is a common regulatory mo-

tif in gene regulatory networks called single-input module (SIM) [155,156]. The SIM network motif is a simple pattern in which one master transcription factor controls a group of target genes. SIMs often occur in sensory transcriptional networks, and they control a group of genes according to the signal sensed by the master transcription factor. In bacteria *Escherichia coli*, the SIM motifs often generate temporal programs of expression in which target genes are turned on one by one in a defined order, and these temporal programs are found to match the functional order of the gene products; the proteins are made just when needed. These proteins work sequentially to synthesise a desired metabolite atom by atom, in a kind of molecular assembly line. This pattern has been observed in the arginine biosynthesis system in *Escherichia coli* [157]. The target genes in a SIM always have a common biological function. Therefore, SIMs often regulate genes that participate in a specific metabolic pathway. Other SIMs control groups of genes that response to a specific stress. These genes produce proteins that repair the different forms of damage caused by the stress. Such stress response systems usually have subgroup of genes that specialize in certain aspects of the response. So if genes coding a set of enzymes along a metabolic or stress-related pathway show similar response to perturbation in gene expression, and there is a TF having high expression before or at the co-expression, it would give evidence of the co-regulation of these enzymes already in transcriptional level by the transcription factor. For more details, see [131, 155].

When studying correlations between gene expression and metabolite concentration levels, the traditional interpretation is that the metabolism is regulated by changes in the gene expression levels. However, the correlations between metabolites and transcripts may often be due to the regulation of transcript levels by metabolic status [120]. An evidence of this comes from a study of *Arabidopsis thaliana* in which diurnal variation of gene expression, metabolite concentration levels and enzyme activities were measured in a wild type, and in a starchless phosphoglucomutase mutant (*pmg*) [158]. The *pmg*-mutant accumulates large amounts of sugars in the light but experiences carbon starvation in the night. The transcript levels undergo marked and rapid changes during diurnal cycles in both genotypes, and are further enhanced in *pmg* compared with the wild type plants. Enzyme activities in both the wild type and the mutant do not show large diurnal changes. Majority of the metabolites show slow changes in the wild type plant. The authors claim that the enzyme activity profile and the metabolite profile change slowly because they represent an integration over time of faster but more transient changes in transcript levels. The slow response of enzyme activities, and the resulting time delay between changes in transcript levels and metabolite levels implies that the correlation between transcripts and metabolites are likely to reflect a regulatory impact of metabolites on gene expression. The amplitudes of the diurnal changes in metabolite levels in the mutant are small, except for sugars in the *pmg*-mutant. The larger diurnal changes in sugar levels in the mutant lead to exaggerated diurnal changes in the levels of more than 4000 transcripts. Many metabolite-transcript correlations were found, and the proportion of transcripts correlated with sugars increased dramatically in the starchless mutant. These results suggests that correlations between metabolites and transcripts may often be due to the regulation of transcript levels by

metabolic status, and are not the result of the regulation of metabolism by changes in gene expression as traditionally interpreted in the literature [120].

2.4.5 Quantification of the metabolome

Two completely different technologies are available for metabolome detection and quantification, namely a mass spectroscopy (MS) and a nuclear magnetic resonance (NMR) spectroscopy [124, 159–161]. In addition to quantifying the proteome, the mass spectroscopic methods are used to determine the *metabolome* of an organism [140, 162, 163]. This thesis will focus on mass spectrometric technologies in metabolic profiling that can be used, for example, to classify healthy and diseased samples, or to understanding the effects of environmental, genetic or other induced changes on organism's metabolism. There are three types of metabolomic analysis: metabolic profiling, metabolic fingerprinting and metabolic footprinting. In metabolic profiling. i.e., targeted metabolomics, quantitative analysis of a predefined set of metabolites is performed. The predefined set can be a selected biochemical pathway or a specific class of compounds, i.e. amino acids. The advantage of the metabolic profiling is the maximization of specificity and sensitivity of the applied methods. The goal in the targeted metabolomics is often to determine metabolite fragmentation patterns and construct calibration curves for absolute metabolite quantification.

On the other hand, metabolic fingerprinting, i.e. untargeted global metabolomics, is a global screening approach aiming to maximize the coverage of analyzed metabolites. Metabolic fingerprinting has been used to classify samples based on metabolite patters that change in response to disease, environmental or genetic perturbations. In metabolic fingerprinting, the goal might just be obtaining a good classifier to classify samples into healthy and diseased classes without further identification or quantification of metabolites. Metabolic fingerprinting can be used to address the success of treatment strategies by monitoring if the metabolic phenotype shifts back to the healthy state after drug treatment. Metabolic fingerprinting does not directly contribute to the biochemical knowledge and understanding of the underlying biological mechanisms because it compromises with the sensitivity and specificity for any particular metabolite. Metabolite fingerprinting requires usually more data analysis, and interpretation of the hundreds to thousands of resulting compounds is challenging due to a large number of unknowns and experimental uncertainties. Hence analysis depends extensively on computational tools, statistical methods and metabolite databases.

The third type of metabolic analysis, metabolic footprinting, denotes an analysis of extra-cellular metabolites in cell culture medium as a reflection of metabolite excretion or uptake by cells. [164–166]. In this thesis, the technique to obtain metabolomics data is an intermediate form of metabolic profiling and metabolic fingerprinting: only certain types of metabolites are investigated but instead of their absolute levels, only relational levels of metabolites are obtained.

Mass spectrometric (MS) technologies are increasingly used methods to identify and quantify metabolites in biological samples [140, 166]. The application of MS to

metabolite analysis is providing new insights into the biochemical functions, and the cellular physiology of living organisms. MS methods have been used for biomarker discovery, and for systematic investigation of metabolic dynamics, for instance. The metabolome represents a vast number of components that belong to a wide variety of compound classes such as amino acids, lipids, organic acids, sugars and nucleotides. These metabolites in living organisms have very diverse physicochemical properties and occur at different abundance levels. There is no single-instrument platform that currently can analyze all metabolites. Furthermore, the comprehensive investigation of the metabolome is being complicated by its enormous complexity and dynamics.

Mass spectrometric methods are based on ionization of the sample metabolites, and accelerating the charged molecules in an electromagnetic field, resulting to the separation of molecules having different mass and charge. Generally mass spectrometry measures the mass-to-charge ratio of the charged molecules, m/z. The m denotes to the mass of an ion and the z is the number of elementary charges on the ion. MS procedure consists of five parts: first, a sample is loaded to the MS instrument, and it is vaporized; second, the sample molecules are ionized by one of a variety of methods, e.g. by bombarding the molecules with an electron beam, which results in the formation of charged particles. Third, the ions are separated according to their mass-to-charge ratio by an electromagnetic field; fourth, the ions are detected, typically using a microchannel plate or photomultiplier tube, and quantified, and fifth, the ion signal is processed into mass spectra. The mass spectrometric data analyzed in this thesis were measured using Thermo Finnigen's LC-LTQ high-resolution MS instrument which ionizes the samples using electrospray ionization, and isolates and fragments the ions using linear ion-trap. Only techniques used to obtain the data are described in this section

Before loading the sample to the MS instrument, different metabolites are usually separated by a chromatographic method such as gas chromatography (GC) or liquid chromatography (LC) [167]. Combination of MS instrument with a separation technique reduces the complexity of the mass spectra due to metabolite separation in a time dimension, provides isobar (different metabolites having the same elemental structure) separation, and delivers additional information about the physicochemical properties of the metabolites. The different metabolites are separated in the chromatographic columns by their different adhesion properties to the column material or the attachment to the small pores. The columns are manufactured from silica with various attached groups. Therefore, the metabolites come out of the column, and enter the MS instrument at different times. The time a solute takes to travel through the column is denoted as a retention time. When using GC-MS technique, the compounds in a sample are dissolved to a mobile gas phase, and separated by a gas chromatograph. The GC-MS method is the most widely used MS method. However, the compounds used in GC-MS need to be volatile and thermostable. In contrast, LC suited better for non-volatile metabolites separates compounds chromatographically using liquid mobile phase, usually a mixture of water and organic solvents. For more information, see [165, 166].

MS requires formation of gas phase ions that can be resolved through the manipulation of electromagnetic field. The ionization method used in this work, the

electrospray ionization (ESI), is suitable for non-volatie molecules separated by LC as it allows desorption and ionization of a wide range of molecules directly from the liquid phase. The ESI is based on the formation and drying of charged liquid droplets. Fine droplets are formed through a charged nebulizer needle. As solvent evaporates, the charges are concentrated on the surface. When surface charge repulsion exceeds surface tension, droplets break up into smaller droplets. This consecutive concentration process continues until gas phase ions are ultimately formed. The ESI produces even-electron pseudo molecular ions $[M+H]^+$ or $[M-H]^-$ usually without fragmentation. Therefore additional fragmentation techniques are required such as tandem MS methods discussed in the subsequent paragraphs [166].

Mass analyzers separate ions according to their mass-to-charge ratio. In order to distinguish between compounds with the same nominal mass, high-resolution mass spectrometers are the instrument of choice. High-mass resolution instruments allow accurate mass measurement, as well as the calculation of empirical molar masses, and thus the elemental formulas of the molecules. There exists several types of mass analyzers with a static or dynamic field. Moreover, the field can be magnetic, electric or both. An example of a mass analyzer is a quadrupole mass filter that uses oscillating electrical fields to selectively stabilize or destabilize the paths of ions passing through a radio frequency quadrupole field created between 4 parallel rods. Only the ions in a certain range of m/z -ratios are passed through the system at any time and detected. The mass analyzer used in this work is the ion trap which is closely related to the quadrupole mass analyzer [168]. In the ion trap, the ions are trapped by a quadrupole and subsequently ejected. The combination of several different mass analyzers can be used to acquire fragmentation information by isolating and fragmenting target ions, and analyzing and detecting the resulting fragments. These are called MS/MS analysers or tandem MS. Fragmentation experiments are important to enable the comparison of experimental fragmentation patterns with authentic standards and spectral databases to confirm molecular structure. For more details, see [165,166]. In addition, combining different MS techniques together increases the amount of detectable metabolites. Platforms that combine technical approaches have covered hundreds of metabolites from plant central and secondary metabolism [169].

Many MS methods require a sample preparation step that is, however, one of the major sources of error in the MS analysis. The steps in the sample preparation are quenching, cell harvesting, cell lysis and metabolite extraction [170]. Quenching to halt the metabolism is usually done through addition of cold methanol or flash-freezing the biological samples in cold nitrogen. Biological matrix consisting of all other biological material than metabolites of interest must be removed from the samples. For metabolite extraction, a wide range of solvent systems and temperatures can be used. All of these steps in the sample preparation can cause metabolite losses. Metabolites may change during the harvest and quenching of the biological material, and they may be lost or modified during the preparation of the extracts. Therefore, the extracts must be converted into a form in which metabolites are stable and that can be applied to the analytic machine. For more information, see [165, 166].

The minimum information about a metabolomics (MIAMET) experiment which

should be reported with each study have been introduced [118]. These are sampling, sample preparation, sample analysis including metabolite separation and MS detection, data export and data analysis. Some reporting standards for metabolomics analysis have been presented [171–173] and the demands on data quality and validation discussed [161].

2.4.6 Statistical analysis of the metabolomics data

After the mass analysis and the detection of the fragmented ions, the metabolites are identified and quantified. The raw data is usually a subject of several normalization and standardization pre-processing methods, for example, to reduce noise and background, to correct the baseline, and to minimize the impact of variability of high-intensity peaks. Mass spectrometric or chromatographic peaks are automatically picked and annotated to certain metabolites. MS instrument manufacturers usually provide a software for the export of MS data such as LECO and Micromass [167]. Several platform-independent metabolomics softwares have also been developed such as XCMS [174], MZmine [175], KNApSAck [176] and msInspect [177]. The data used in this thesis was collected using Xcalibur-program. More information about the form of data produced by MS can be found in [165, 166].

The absolute values of metabolite concentrations can be obtained when inner standards with known concentrations are used to calibrate the MS instrument [165, 166]. Sometimes, only few inner standards not necessary including the target metabolites can be used to determine the relational levels of metabolites as was done to obtain the data analyzed in this thesis.

Statistical analysis methods usually applied to the data matrices obtained from the metabolome analysis are the same as for the gene expression data: principal component analysis (PCA), independent component analysis (ICA) [178], hierarchical clustering, self-organizing maps (SOM), and canonical correlation analysis (CCA). These unsupervised methods suit for a situation in which the aim is to cluster samples of unknown identity, and investigate the metabolites responsible for the cluster differences. On the other hand, the sample identity is often known, for example, when the aim is to discover characteristic biomarkers. Then supervised methods, such as analysis of variance (ANOVA) [179], or partial least squares methods (PLS), can be applied. PLS methods identify the combinations and weightings of a large set of metabolites that provide the best prediction of a target trait such as biomass production [180]. One type of PLS method, O2PLS [21, 181, 182] allows systematic predictive variation that links different types of parameters to be separated from parameter-specific variation. Examples of applying statistical and modelling methods to metabolomic data on its own or together with other data types can be found in [158, 171, 171, 180, 183–192].

Several sources produce variation to the metabolomic data: First, specific perturbations within the underlying network of biochemical reactions, typically an overexpression or knock-out of a gene coding for an enzyme, causes changes in metabolite levels. Second, global perturbations, such as a circadian clock, and responses to temperature and stress induce changes at multiple sites within the metabolic net-

work. Global changes that are brought about by external factors can influence a large number of metabolites simultaneously. Third, there is intrinsic variability between strongly interrelated variables. Popular method to analyse metabolomics data is to study the correlations between the metabolites to find functionally related metabolites [193–196]. Correlations are transitive, meaning that given a metabolite A which shows a high correlations to metabolite B, as well as to metabolite C, then B and C should be correlate as well. However, correlations do not tell much about the causal relationships between metabolites.

Metabolites are correlated due to various reasons [196] such as equilibrium between neighbouring metabolites; mass conservation; diminutive fluctuations within the metabolic network which propagate through the system [197–199]; global perturbations such as the circadian rhythm; asymmetric control in which one parameters dominates the control of several metabolites; shared biosynthetic enzymes; and coordinated regulation of gene expression [21, 197, 198, 200]. Whether two metabolites correlate or not is a combined result of many if not all, biochemical reactions, regulatory interactions and the inducing fluctuations that constitute a system. Correlations do not necessarily occur between metabolites that are neighbours on the metabolic map but other more subtle mechanisms are involved [197, 198]. Therefore, many correlations observed in metabolome data remain unclear, and additional knowledge than just metabolite and transcript levels is needed to decode the regulatory mechanisms of metabolic pathways hidden in the metabolite-metabolite correlations [186, 201].

Time series data is essential when studying correlations between metabolites and to draw conclusions about the regulatory processes [179, 202]. Time series experiments can reveal temporal dependencies between parts of metabolite network and provide insights into functional dependencies between the metabolites. Further insight brings the combined analysis of gene expression and metabolites. For example, an integrated time series metabolome and transcriptome analysis on a mutant plant of abscisic acid (ABA) biosynthesis under dehydration has been performed. This study indicated the ABA-dependent regulation for the formation of branched-chain amino acids (sacharopine, proline, polyamines) which presumably play important roles in plants' responses to dehydration stress [203].

2.5 Regulation of oxidative stress signalling in *Arabidopsis thaliana*

2.5.1 Oxidative stress signalling

Plants are the most consummate and sophisticated chemical systems in the world. They use light energy to convert carbon dioxide into carbohydrates in their leaves. However, plants cannot move and so they need to cope with chancing environmental conditions and various biotic stresses including pathogen attacks, and abiotic stresses such as heat, drought, air pollutants and absence of nutrients.

Ozone is a triatomic oxygen molecule protecting life on earth from ultraviolet radiation in stratosphere. However, when appearing in troposphere, it is an

atmospheric pollutant and causes oxidative stress to plants. The tropospheric concentration of ozone has increased two- to five-fold during the past 60 years due to human actions [204]. High peaks in ozone concentration cause visible damage in sensitive plants by oxidizing and damaging cell membranes [205–207]. During a continuous low level of ozone, photosynthesis and growth of a plant are impaired, and the senescence occurs earlier, which results in crop losses worth of several million dollars annually [207–209]. Therefore, is it important to understand the signalling during oxidative stress, for example, to develop stress-tolerant genetically modified (GMO) crops.

The events starting from ozone recognition, and resulting in adaptation of a plant to a stressful situation or tissue damage are tightly regulated. This oxidative stress signalling resembles the hypersensitivity response (HR) of plants to pathogens such as bacteria and viruses [206,210,211]. HR is a form of programmed cell death (PCD), apoptosis. Apoptosis requires active cellular processes such as energy production, signal transduction, ion fluxes, transcription and translation [212].

Under intolerable concentrations of ozone, the oxidative stress signalling leads to the PCD and visible tissue damage, lesions. The oxidative stress signalling and regulation of the apoptotic lesion formation have six stages: regulation of ozone flux to leaves; ozone degradation to reactive oxygen species, and detoxification by ascorbate in the apoplast; ozone sensing or perception; ozone-induced active production of reactive oxygen species; early activation of mitogen-activated protein kinases by ozone; and hormonal regulation of ozone-induced lesion formation. These components are presented in Figure 3, and the steps 1 to 11 in this figure are explained in the subsequent paragraphs.

The plant cells are surrounded by a free diffusional space called apoplast. It is formed by the cell walls of adjacent cells, as well as the extracellular space. The apoplast facilitates the transport of water and solutes across a plant tissue or an organ. Ozone enters to the plant through small pores in leaves which are used by the plant to exchange gases and to evaporate water with the apoplastic space and air. The mechanism by which the plant cells sense ozone is not known. Ozone itself is not deleterious to the cell plasma membrane, but when it enters to the apoplastic fluid, it rapidly degrades into superoxide anion radicals O_2^-, hydrogen peroxide H_2O_2, hydroxyl radicals $OH^\bullet$, and singlet oxygen. These are called reactive oxygen species, ROS (step 1 in Figure 3). Although ROS damage the cellular membranes [207], they are also important regulatory molecules in several stress signalling processes. During a substantial or prolonged ozone exposure, ROS signalling eventually initiates a programmed cell death of individual cells, or the whole tissue [211]. In apoplast, the ROS molecules are scavenged by apoplastic antioxidative molecules such as ascorbate. This partly removes the harmful ROS generated from ozone (step 1 in Figure 3) [214]. If the apoplastic antioxidants cannot detoxify ROS, ROS level in apoplast increases over a threshold which elicits several downstream processes [213].

If apoplastic antioxidative capacity is not enough to remove the ROS formed from ozone, the perception of ROS induces a signalling cascade that include endogenous signalling, as well as cell-to-cell signalling. Different mechanisms have been proposed for ROS sensing [213, 215–217] but the whole picture has not emerged. When ROS

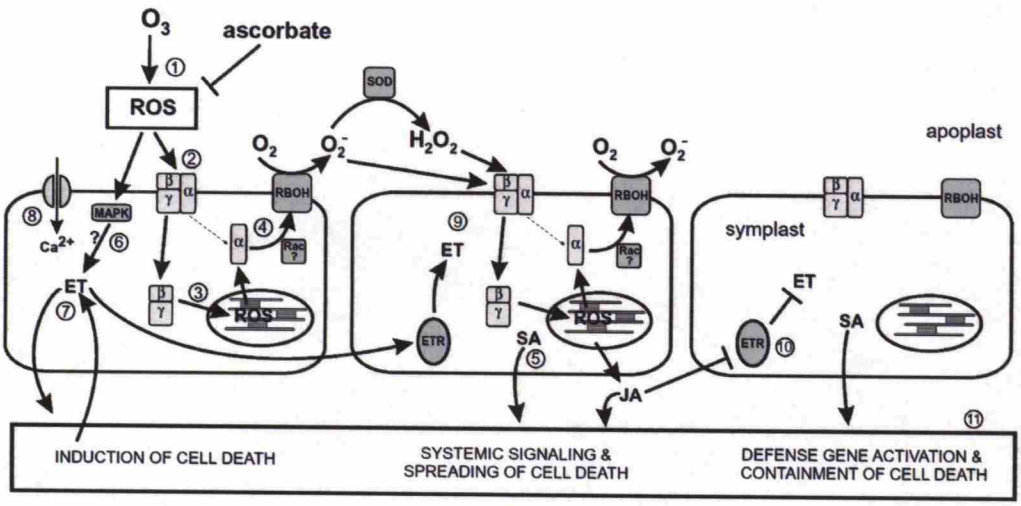


Figure 3: ROS signalling in oxidative stress. The rounded rectangles are individual plant cells. ROS, reactive oxygen species; O_3, ozone; MAPK, mitogen-activated protein kinase; Ca^{2+}, calcium-ion; ET, ethylene; α, G-protein α-subunit; β, G-protein β-subunit; γ, G-protein γ-subunit; Rac, a small GTPase; RBOH, NADH oxidase (respiratory burst oxidase homolog); SOD, superoxide dismutase; O_2, oxygen; O_2^-, superoxide radical; H_2O_2, hydrogen peroxide; ETR, ethylene receptor; SA, salicylic acid; JA, jasmonate. [213]

are perceived by a cell, the cell triggers a signalling pathway which leads the cell to produce extracellular ROS. The extracellular ROS are then perceived by the neighbouring cells. The neighbouring cells also start an active production of ROS, and thus the signal propagates. Various protein-signalling cascades are involved in this process, as well as several plant hormones. Hormones regulate especially the ROS signalling initiation, propagation and containment. This signalling can lead to a programmed cell death. The place where the PCD signalling starts is called lesion initiation site.

Various signalling proteins such as guanosine nucleotide-binding proteins (G-proteins), NADPH oxidases and mitogen-activated protein kinases (MAP kinases) are likely involved in the ozone stress signalling (steps 2, 3, 4 and 6 in Figure 3) [218–220]. There are two separate chains of events that occur when the ROS are formed in the apoplastic space. First, somehow ROS production is induced in the chloroplasts through the heterotrimeric G-protein (or the $G\beta\gamma$ complex) (steps 2 and 3 in Figure 3) [221]. In addition, as a result of chloroplastic regulation, there is a rapid and coordinated down-regulation of many nuclear genes encoding chloroplast proteins, presumably as a result of chloroplast-derived signals [222]. Second, ROS production in the cell plasma membrane by NADPH oxidases (RBOH in Figure 3) is activated in a G-protein α-subunit-dependent way, possibly through a small GTPase Rac (step 4 in Figure 3). For more details, see [213].

Receptor enzymes receiving signals from extracellular and intracellular space usually initiate a long series of protein-signalling cascades which eventually lead to

expression of certain genes as a response to the signal. In these cascades, the signal usually propagates from protein to protein by phosphorylation. MAP kinase cascades are an example of such signal transduction. MAP kinase cascades are activated within minutes from the beginning of ozone exposure (step 6 in Figure 3). The activation of MAP kinases does not require the G-protein, and they do not appear to be directly involved in the activation of ROS production by NADPH oxidase RBOH. The two primary oxidative stress related MAP kinases in Arabidopsis are AtMPK6 and AtMPK3 marked in Figure 4. The Figure 4 shows also the upstream MAP kinase kinases, and MAP kinase kinase kinases, and downstream targets of which some are transcription factors. The MAP kinases AtMPK6 and AtMPK3 are connected to the hormonal responses, especially to the synthesis of plant hormone ethylene [213]. The activation of AtMPK6 might results in increased plant hormone ethylene synthesis (step 6 in Figure 3 and Figure 4). In addition to ethylene, a biosynthesis of another plant hormone, namely salicylic acid (SA) is also induced in the beginning of ROS signalling, and it accumulates during lesion spreading (step 5 in Figure 3) [223, 224]. Salicylic acid is also involved in the PCD signalling and contributes to the containment of lesion growth [213]. ROS has been shown to be involved in the regulation of key steps in SA biosynthesis during pathogen infection.

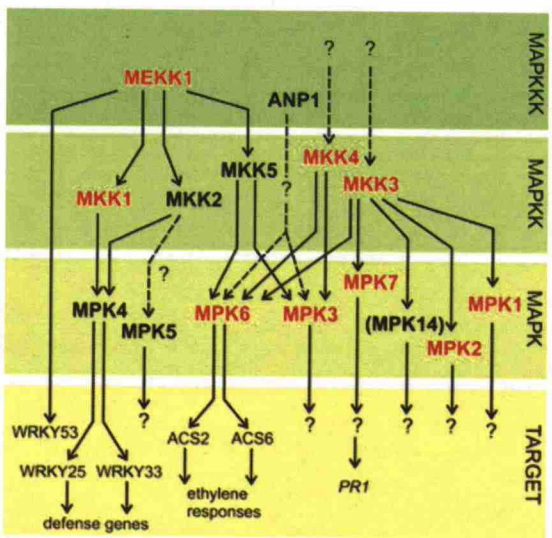


Figure 4: MAP kinase cascades and some transcription factors and metabolic pathway genes which are affected by them. MAPK, mitogen-activated protein kinase; MAPKK, MAP kinase kinase; MAPKKK, MAP kinase kinase kinase; PR-1, pathogen response-1; ACS, 1-aminocyclopropane-1-carboxylic acid synthase, the rate-limiting enzyme in ethylene synthesis. [219]

The activation of MAP kinases and NADPH oxidases are both calcium dependent, and indeed calcium influx takes place in the early stages of ROS signal transduction (step 8 in Figure 3). After step 4, when the ozone has been sensed, and the G-protein α-subunit has activated the NADPH oxidase (RBOH) in the cell membrane, the NADPH oxidase produces extracellular ROS to the apoplastic space, and ROS production and initiation of the SA-dependent cell death spread to the adjoin-

ing cells. This active production of ROS is called an oxidative burst [224–227]. In the oxidative burst, apoplastic O_2^- and H_2O_2 are produced. Thus, NADPH oxidase passes the signal from cell to cell. The ROS produced by the NADPH oxidase in the lesion initiation site are perceived by the neighbouring cells, in which G-protein controlled activation of ROS-production is induced in several subcellular compartments such as chloroplast and mitochondrion.

Hormonal regulation and plant hormones abscisic acid (ABA), salicylic acid (SA), ethylene (ET) and jasmonate (JA) are important in determining the degree of plant ozone sensitivity and lesion initiation, propagation and containment [212, 228–232]. Increase of ethylene concentration is one of the fastest responses of plants to ozone (step 7 in Figure 3), and is well correlated with the extent of cell death [231, 233–238]. Furthermore, ROS production drives the SA-dependent cell death (step 5 in Figure 3) [223, 239] which continues until the third hormone, JA, antagonistic to lesion propagation, contains the enlargement of the lesion lesions spread (step 10 in Figure 3). The spread of the ROS signal to the neighbouring cells, and the competence of these cells to receive the ROS signal is ethylene dependent (steps 7 and 9 in Figure 3). The heterotrimeric G-proteins are involved in receiving the signal, and they pass it forward in an ethylene dependent manner (step 9 in Figure 3). The repetition of this process along cells is depicted as a self-amplifying loop, i.e an oxidative cell cycle, presented in Figure 5, which results in the formation of visible lesions [231, 232, 240]. In this cycle, the endogenous, cell-death driving ROS production triggered by ozone is ET and SA dependent and repressed by JA [205, 231, 236, 237]. Cell death during the lesion spread creates an increasing concentration of membrane lipid peroxidation products which serve as a substrate for jasmonate (JA) biosynthesis. Therefore, when the lesion propagates, jasmonates begin to accumulate. Increased JA accumulation reduces ethylene sensitivity of the cells in the spreading lesion, resulting in a gradual decrease in the ethylene-dependent ROS production, and consequently containments the cell death (step 10 in Figure 3). The JA signalling is suppressed by SA and ET during the initial cell death. Abscisic acid (ABA) is still another plant hormone functioning in the oxidative stress signalling. The biosynthesis of ABA is increased by ET. Moreover, the balance between ABA and ET is maintained through subsequent inhibition of ET biosynthesis by ABA and by mutually antagonistic interactions between ABA and ET signalling. In addition, ABA is involved in responses to several abiotic stresses, and is a regulator of glucose signalling. For more details, see [213].

All steps from 1 to 11 in Figure 3 regulate a set of ozone-induced responses including cell death, and the transcription of defence-related genes (step 11). As a conclusion, the time development and activation of each process in oxidative stress signalling starting from the ozone exposure is shown in Figure 6 [213].

A major part of plant adaptation to abiotic stress is regulated at the level of transcription. Several processes contribute to the stress signalling: post-translational activation [241]; selective nuclear import of transcription factors; regulation of DNA accessibility by chromatin modifying, and regulation of remodelling enzymes; and cooperation between two or more TFs in a stress-responsive promoter. Several hundred to thousands of genes in plants have been identified with altered response to abiotic

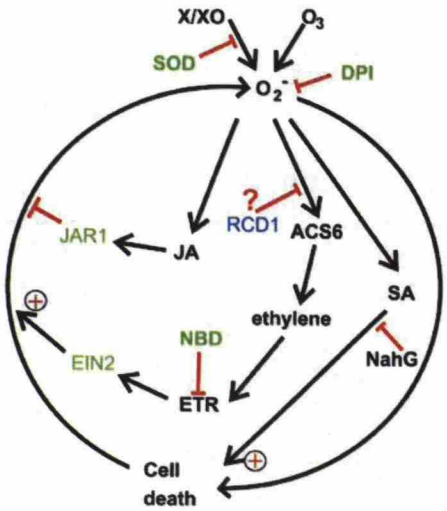


Figure 5: Oxidative cell cycle. X/XO, xantine/xantine oxidase(oxidative agent); SOD, superoxide dismutase; O_2^-, superoxide radical; O_3, ozone; DPI, diphenylene iodonium(an inhibitor of ROS accumulation) ; RCD1, RADICAL INDUCED CELL DEATH; JA, jasmonate; JAR1, mutant of this gene is JA insensitive ; ACS6, ACC synthase 6; ETR, ethylene receptor; NBD, ethylene antagonist; EIN2, mutant of this gene is ethylene insensitive; NahG, salicylate hydroxylase; SA, salicylic acid. [231]

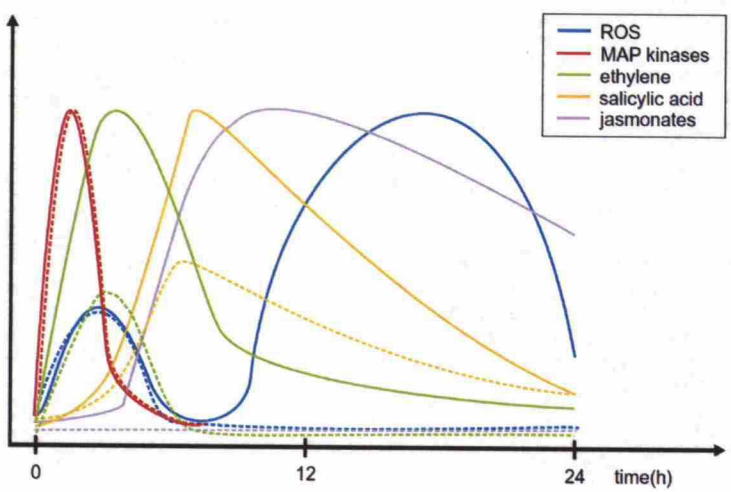


Figure 6: Time development of oxidative stress. [213]

stress [242]. To translate the stress exposure into appropriate changes in gene expression, suitable signalling pathways needs to be activated, ultimately ending up with a transcription factor binding at the promoter of the gene to be transcribed. Changes in gene expression can be detected within 15-30 min of an applied stress [243] and can last for several days. The expressed genes can be classified into three groups: early (expressed at 0-3 h after the onset), intermediate (expressed at 3-6 h after

onset) and late genes (6-10 h or more after the onset). Examples of genes induced early are coding pathogen-responsive proteins and an elicitor-induced genes, examples of the intermediate genes are members of the phenylpropanoid pathway, and the late genes include peroxidases and hydroxyproline-rich glycoprotein genes [206].

Arabidopsis genome encodes about 2000 TFs, corresponding to 7 % of the genes in this species [244]. The function of about ten of them is known. TFs have an important role in oxidative stress signalling. In response to oxidative stress signalling, some TFs already present in the cytoplasm can be activated by post-transcriptional modifications, or they can be transferred to the nucleus. On the other hand, some TF genes are transcribed to produce new regulatory proteins. Many genes with rapid changes in gene expression after abiotic stress treatment encode TFs [243,245]. TFs can be divided into early and late expressed. The early TFs increase the expression of a second class of TFs with a role in later and prolonged stress responses. For more details, see [246].

In this thesis, the metabolite data contain 19 proteinogenic amino acids: alanine, arginine, asparagine, aspartic acid, glutamate, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, valine; 1 polyamine: γ-aminobuturic acid (GABA), 6 non-proteinogenic α-amino acids: citrulline, ornithine, homocysteine, cystathionine, α-aminoadipic acid, α-aminobutyric acid; and 2 other organic acids with an amino group: β-aminoisobutyric acid and pipecolic acid. Most of these metabolites are produced in the chloroplasts where the photosynthesis occurs. In addition to being precursors for protein synthesis (proteinogenic amino acids) and secondary metabolites, these metabolites have several other functions in a cell, for example in oxidative stress regulation [206].

Chloroplasts-derived metabolites, and especially amino acids, serve as precursors for biosynthesis of several plant hormones and other biomolecules that regulate the oxidative stress response: cysteine is a precursor of glutathionine (GSH); methionine is a precursors of ethylene; phenylalanine is a precursor of SA; tryptophan is a precursor of the growth hormone auxin and abscisic acid; arginine is precursor of agmatine, a polyamine, whereas ornithine is a precursor of polyamines putrescine, spermidine and spermine; and glutamate is a precursor of polyamine gamma-aminobutyric acid.

Aromatic amino acids serving as precursors for the biosynthesis of auxin and flavonoids are synthesised in chloroplast via a shikimic acid pathway [247–249]. Besides providing building blocks for protein synthesis, aromatic amino acids also serve as precursors for important secondary metabolites, including tryptophan-derived indole hormones and phenylalanine- and tyrosine-derived flavonoids and lignins (Figure 7) [247, 248]. The shikimate-derived pathways are regulated by complex networks, including feedback control by aromatic amino acids and environmental factors [249]. A proteinogenic, aromatic amino acid phenylalanine is a starting compound in the flavonoid biosynthesis. Moreover, together with tyrosine derived from phenylalanine, phenylalanine is the precursor of lignin (Figure 7). Ozone exposure causes increases in the activities of several phenylpropanoid and flavonoid pathway enzymes [206]. Ozone treatment has been shown to increase the activities of PAL and CAD, enzymes controlling respectively the phenylpropanoid, flavonoid and

lignin biosynthesis pathways (Figure 7). These pathways play a significant role in plant defence responses because they synthesise many potentially protective compounds including flavonoids, furanocoumarins and lignin. For more details, see [206].

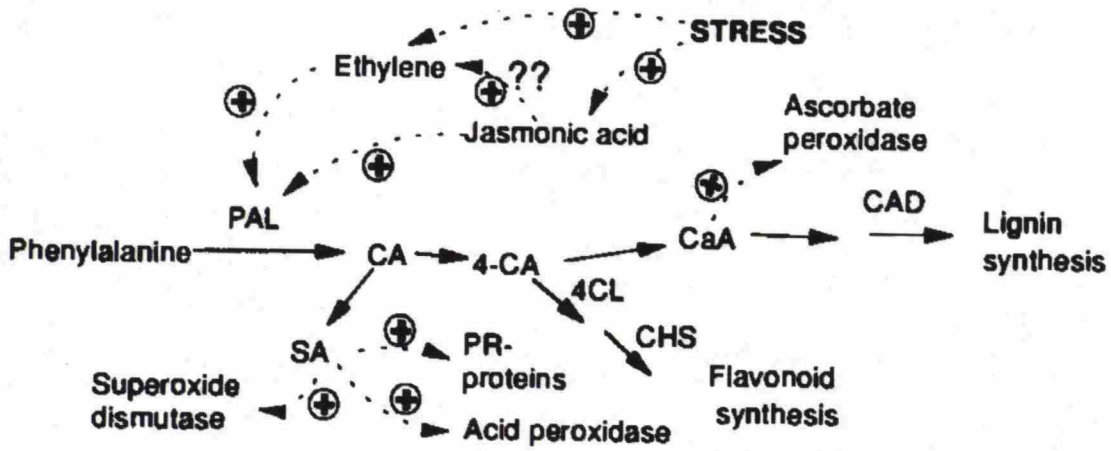


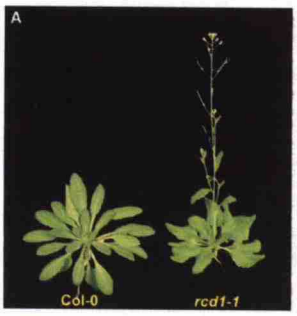
Figure 7: Connections between hormones, secondary metabolites and phenylalanine. This graph depicts the stress induction of phenylpropanoid metabolism in plants, and induction of various defence reactions by some phenylpropanoid substances. Dashed lines indicate known induction of enzyme activity or gene expression. CA, cinnamic acid; 4-CA, trans cumarylic acid; CaA, cafeic acid; SA, salicylic acid; CHS, chalcone synthase, PAL, phenylalanine-ammonium lyase; CAD, cinnamyl alcohol dehudrogenase, 4CL, p-coumarate-CoA ligase. [206]

This biological background to the oxidative stress signalling is given to present connections between measured metabolites, and other components participating in the signalling. It was recognized that several of the measured metabolites are precursors of hormones, or other metabolites functioning in the signalling or stress adaptation. The results obtained by this thesis are compared to the known regulation of oxidative stress signalling presented in this chapter. The chapter also illustrates the importance of transcriptional and metabolic regulation during the oxidative stress signalling.

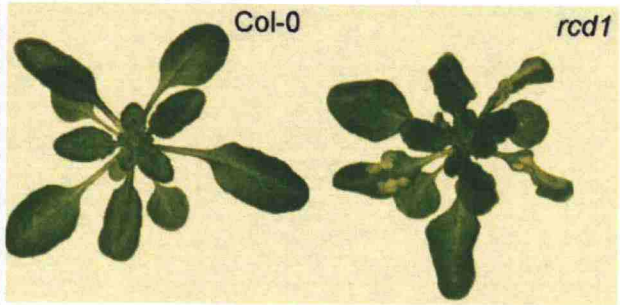
2.5.2 Role of the gene *RCD1* in oxidative stress signalling

The protein RCD1 (RADICAL-INDUCED CELL DEATH1) mapped to a gene AT1G32230 has an unknown role in oxidative stress signalling. This protein is a member of SRO (Similar to RCD-One) protein family, consisting of 6 members, which all contain a certain protein-protein interaction domain. An ozone tolerant *Arabidopsis thaliana* ecotype Col-0, when having a mutation in this gene, gives an ozone-sensitive phenotype. Mutation disrupts an intron splice site resulting in a truncated protein. In addition to ozone sensitivity, the mutant *rcd1* has several other phenotypes related to growth, development and hormone biology [231, 250].

The *rcd1* is sensitive to condition that leads to the formation of apoplastic ROS but it has an increased tolerance to UV-B light [251]. Moreover, the appearance of the mutant plant is different; it has smaller, more erect rosettes, altered leaf shape and earlier flowering (Figures 8(a) and 8(b)).



(a) Phenotype differences of the wild type Col-0 and *rcd1*. [231, 250]



(b) The wild type Col-0 and the *rcd1* mutant after ozone exposure. The Col-0 shows no sign of damage, whereas the *rcd1* mutant has produced lesions. [231]

Figure 8: *Arabidopsis thaliana* wild type Columbia-0 and *rcd1* mutant.

The RCD1 is assumed to interact with another proteins, and likely with transcription factors because the RCD1 protein contains several protein-protein interaction domains [252], and domains which localize it into the nucleus. [252]. The protein-protein interaction domain in the RCD1 was shown to interact with several transcription factors or DNA-binding proteins involved in dehydration or osmotic stress responses, such as DREB2A, which is a central transcription factor in the ABA-independent responses to osmotic stress [253]. Indeed, one of the target genes of DREB2A, namely the RD29A/LT18, had lower steady state transcript levels in the *rcd1* mutant. It has also been speculated that the RCD1 could function by post-transcriptionally modifying its target proteins [213].

The *rcd1* is a lesion propagation mutant and hormone-balance -related runaway mutant. Its sensitivity to extracellular ROS leads to extensive formation of HR-like lesions, and activation of several pathogen defence-related processes in response to ROS. Therefore, the *RCD1* seems to define a component of the HR related to the role of ROS in PCD regulation. The *rcd1* appears to be deficient in the control of lesion propagation, leading to wide cell-death spreading, whereas in the same conditions, the wild type has no visible lesions, although microscopically small lesions occur (Figures 8(a) and 8(b)). The deficiency occurs only transiently in the *rcd1* because after 24 h (ozone exposure between 0 h and 6 h) the propagation of lesion is brought under control. In contrast, in the wild type Col-0, there is lesion initiation but no propagation.

The mutation in the *RCD1* seems to affect mostly hormone-related processes. The *rcd1* has normal levels of SA and JA under control condition but under ozone exposure, the levels of ET, SA and JA are higher compared to the wild type. In addition, in a microarray analysis with 6400 oxidative stress related genes, the few genes that had changed basal expression level in clean-air grown *rcd1* are involved in

either ethylene-, ABA- or sugar-related processes [250]. Furthermore, the *rcd1* shows slight insensitivity to ABA, ET and JA in some specific responses to these hormones. The *rcd1* is described as an ET overproducer and a JA insensitive mutant. It has been hypothesized that the *RCD1* encodes a protein that most likely is involved in interactions between hormonal signalling cascades in abiotic stresses [213, 250].

While the *rcd1* mutant shows differences in hormone levels and metabolism-related processes (sugar metabolism [250]), and some of the metabolites analyzed in this work are precursors of the hormones playing a role in the oxidative stress signalling, changes in metabolite concentrations between the wild type and the mutant *rcd1* under ozone exposure are expected. The results obtained in this thesis are investigated in the light of previous knowledge of changes caused by the mutated gene during the oxidative stress signalling. The mutated gene very likely interacts with the transcription factors, which makes the interpretation of the results hard. If under ozone exposure there is a differential expression of some genes (either genes coding metabolic enzymes or transcription factors) between genotypes, this implies that either the RCD1 directly bounds to the promoter of these genes and regulates them, or the RCD1 interacts with the transcription factors regulating these genes. In the latter case, one could be interested of the potential motifs at the promoters of the differentially expressed genes to find transcription factors which potentially interact with the RCD1.

3 Probabilistic modelling and mathematical background

3.1 Probabilistic modelling

Data analysis is a field of study using computational methods to analyse collections of data including data obtained by measuring natural phenomena. Natural data contain always errors and unwanted variation resulting from sources such as inaccurate measurement technologies, differences in observed subjects, and human errors. The measured phenomenon can also be stochastic by nature. Due to these different sources of random errors in the observed data, the exact underlying description of the phenomenon is unknown. Nevertheless, the underlying model for the phenomenon can still be estimated, and a deviance of the estimate from the true model determined using probabilistic modelling framework. To understand the probabilistic model estimation goal, definitions for some key concepts of statistical modelling are presented in this and the following chapters.

A data set is a collection of n observations or samples x each of which is considered to be an instance of a random variable X. The noise in the individual samples is usually considered independent between samples, and to follow the same distribution if the samples are observations replicated under the same experimental conditions. The samples having this property are referred to as independent and identically distributed (i.i.d.). Each sample can be a continuous or a discrete single value, or a data vector of several values; in that case the instance of a random variable, $\boldsymbol{x}$, is a multidimensional variable with dimension d. Elements in a data vector $\boldsymbol{x}$ are called features. A data set can be represented as a data matrix $\boldsymbol{X}$ in which columns correspond to samples and rows correspond to features. In this thesis, two multivariate data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ with paired samples are analyzed. Paired samples of the two data sets are co-occurring, i.e. the ith column in data set $\boldsymbol{X}$ is paired with the ith column in data set $\boldsymbol{Y}$. Both data sets have an equal number of n samples.

Collections of observations can be described using probabilistic models. A model describing the data enables the understanding of the underlying phenomenon, and the prediction of future observations. A model is usually constructed to describe and summarize the process that generated the data; such models are called generative models. A model is usually a specific instance of a model family that is a collection of all possible models. The model family is usually determined by the assumptions made before the model implementation. Assumptions include, for example, the form of the probability distribution, and the independency assumptions of the variables.

A specific model within a model family is defined by model parameters. A single model is found by estimating the values of the model parameters from the data. Parameter values are chosen by optimizing the data description of the model. This is called learning or fitting a model. To find the best model, a measure for the quality of the model describing the data set is required. The criterion of fit or cost function usually used is the likelihood function. The likelihood function defines the probability of the observed data given the model parameters. By maximizing the

likelihood function, a maximum likelihood estimate for the model is obtained.

A good probabilistic model generalizes well to yet unseen data; it predicts future observations well, and describes the underlying probability distribution and not just the data used in the learning phase. When the model does not generalize to new observations, overlearning or overfitting has occurred. This implies that the model does not only fit to the training data but also models the noise in the training data, which is undesirable. Overfitting is likely when the model is complex, i.e. has high effective dimensionality of parameters, and when the training data set is small. One solution to overfitting is regularization which forces the complex model to a simpler one. Regularization can imply reduction of non-linear models towards linear ones, or drawing certain model parameters to zero. Regularization is often done by adding a separate penalty term to the cost function. Regularization is discussed more detailed in Chapters 3.2, 3.3 and 3.5.5.

Machine learning is a field of study in which computational programs are implemented to learn from data [254]. In machine learning, learning methods are roughly divided into two groups: supervised learning and unsupervised learning. This thesis considers unsupervised learning where the task is to summarize the data, and find the probability distribution that represents the distribution of the data. Examples of unsupervised learning methods are clustering, density estimation, visualization and dimensionality reduction. Exploratory data analysis is a subfield of unsupervised data analysis. In exploratory data analysis there is often little prior information of the data, and the goal is to extract novel systematic properties from an unknown data collection. Properties of the data that are studied by exploratory data analysis can be complex and largely unknown. Exploratory data analysis can suggest novel hypotheses for statistical testing, or suggest what kind of data to collect in the future to test these hypotheses. Examples of exploratory data analysis include clustering and visualization.

Data fusion is a subfield of data analysis aiming at combining several data sources, in order to improve the accuracy of the analysis. Noise in measurements is usually reduced by averaging over multiple measurements; thus data fusion can be seen as a generalization of averaging over several sets of measurements. In addition, the main interest in data analysis can be to find structure visible in several data sets with co-occurring observations. In machine learning, multi-view learning is a term including methods, for example, which search consensus between the multiple views, i.e. multiple data sets. Consensus can be defined in a numerous ways, for example, by seeking the variation common to both data sets. A rationale behind multi-view learning is that if models learned based on multiple views agree with each other, they are likely to generalize better.

In data fusion, co-occurring data sets are thought to supervise each other. Supervision is used to focus fundamentally unsupervised data fusion methods on more relevant findings which in this case are the structure shared in the paired data sets. Supervision is provided, for example, through searching for statistical dependencies between co-occurring data sets. In Chapter 3.4 statistical dependency between two or more data sets is defined and studied more detailed.

This thesis discusses the analysis of high-throughput bioinformatics data sets.

Data analysis in bioinformatics is challenging in many ways: Due to the development of measurement techniques, one can measure the gene expression of thousands of genes at once. However, the data are very noisy, and often the number of samples, n, is much smaller than the number of features, p, such as genes. This is called large p, small n problem [255]. Large number of features compared to the number of samples often impedes the statistical analysis by leading to computational issues, inaccurate estimates of the model parameters and results which do no generalize. However, when obtaining biological high-throughput measurements, thousands of samples from individual humans, animals or other organisms cannot be obtained because there are not enough human patients, or it would be too expensive and ethically questionable if thousands of laboratory animals were used. In addition to the cost of gaining large data sets, there is also cost of learning from large data sets. Moreover, the relationships between variables in biological data can be very complex. The large number of variables, especially when they are not independent of each other and form highly correlated groups, i.e. the variables are multicollinear [256,257], the statistical analysis and biological interpretation of the results becomes difficult [38]. The large p, small n problem, and the overfitting of the models to a data set having the properties mentioned above can be overcome by regularization methods, and by finding a low dimensional representation of the data [255].

Nowadays in machine learning and bioinformatics, it is very popular to fuse several data sets. In this thesis, the data fusion of gene expression and metabolomics data is studied. Hence, the learning scenario is multi-view learning discussed above. Multi-view learning is one way to face the large p, small n problem as fusing several data sources can be seen as averaging over them. Integrating different types of omics data might answer, for example, to the following questions: "Which variables from both types of datasets are related to each other?", and "Which variables are relevant and provide more insight into the biological experimental hypotheses?"

3.2 Regularization in statistics

The high dimensionality and small sample size in a data set cause several problems to the statistical analysis. When the data set is complex due to the high dimensionality, it is tempting to fit complex models with large number of parameters to the data. This leads often to overfitting; the observations are modelled perfectly in the training data but the fitted model is ambiguous: there are several sets of model parameters that explain the data equally well. Overfitted model is also incapable predicting future observations. The large p, small n setting causes the model parameters learned from the data to be unstable, implying that their variance is very high. Therefore, more robust ways to estimate model parameters are needed [255].

Statistical regularization is a class of methods that modify the maximum likelihood of a model to give reasonable answers when overfitting and instability issues are very likely, for example, in a large p, small n situation. Regularization is used to control the model complexity, and to improve the parameter estimation and the generalization to future observations. Regularization can be seen as incorporating prior information to data analysis to better understand the data, and to make the estima-

tion of the parameters robust. Regularization methods often modify the maximum likelihood function by introducing additional penalty terms to it. Regularization methods are also called shrinking, smoothing and penalizing methods. For more details, see [258].

Regularization in linear regression was first introduced by Hoerl and Kennard [24, 259]. A standard model for a multiple linear regression is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\varepsilon} \; . \tag{2}$$

In Equation 2, $\boldsymbol{X}$ is $n \times (p+1)$ data matrix of rank $p+1$, and $\boldsymbol{w}$ is an unknown vector of regression weights of length (p+1). The elements of the regression weight vector are marked as $w_0, w_1, \ldots, w_p$; w_0 is the intercept term, so the first column of the matrix $\boldsymbol{X}$ consists of ones. In Equation 2, $\boldsymbol{\varepsilon}$ is a Gaussian error term having zero mean and precision β (or variance β^{-1}). The maximum likelihood, or the ordinary least squares estimate for $\boldsymbol{w}$ is

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \; . \tag{3}$$

In Equation 3, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ needs to be invertible, i.e. nonsingular. If the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible, the equation 3 does not have a solution, or the solution is not unique. Moreover, the estimate $\hat{\boldsymbol{w}}$ is already unstable, i.e. having high variance, when the matrix $\boldsymbol{X}^T\boldsymbol{X}$ deviates from the unit matrix. Considering that the sample covariance matrix for a zero-mean data is $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$, n being the sample size, the deviation of $\boldsymbol{X}^T\boldsymbol{X}$ from the unit matrix implies that the variables in the data set correlate with each other, i.e. data is not orthogonal and the variables are multicollinear. If the multicollinearity is considerably high, or the number of variables exceeds the number of samples in the data, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ becomes singular. Even matrices close of being singular, i.e. ill-conditioned, cause problems. To demonstrate the effect of singularity and ill-condition of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ to the ML-estimate of $\boldsymbol{w}$, the covariance matrix Cov, $(\hat{\boldsymbol{w}})$, and the mean squared error $\mathbb{E}[L_1^2]$ of $\hat{\boldsymbol{w}}$ are considered:

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{w}}) &= \beta^{-1}(\boldsymbol{X}^T\boldsymbol{X}) & (4) \\ \mathbb{E}[L_1^2] &= \mathbb{E}[(\hat{\boldsymbol{w}} - \boldsymbol{w})^T(\hat{\boldsymbol{w}} - \boldsymbol{w})] = \beta^{-1}\mathrm{Trace}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \; . & (5) \end{aligned}$$

For the mean squared error of the $\hat{\boldsymbol{w}}$ shown in Equation 5, i.e. $\mathbb{E}[L_1^2]$, the following inequality holds:

$$\mathbb{E}[L_1^2] > \beta^{-1}\frac{1}{\lambda_{\min}(\boldsymbol{X}^T\boldsymbol{X})} \; . \tag{6}$$

In Equation 6, $\lambda_{\min}(\boldsymbol{X}^T\boldsymbol{X})$ is the smallest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$. When $\boldsymbol{X}^T\boldsymbol{X}$ deviates from the unit matrix to ill-conditioned or even singular, non-invertible matrix, $\boldsymbol{X}^T\boldsymbol{X}$ has smaller and smaller eigenvalues, and the mean square error for $\hat{\boldsymbol{w}}$ becomes large. In that case, the least squares estimate for $\hat{\boldsymbol{w}}$ becomes unreliable and sensitive to small changes in the data. For more details, see [24, 259].

To solve the instability issue of the regression weight vector, Hoerl and Kennard [24] developed a regularization method called a ridge regression to obtain more

robust estimates. In ridge regression, the maximum likelihood estimate of the regression weights is

$$\hat{\boldsymbol{w}}^* = (\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y};\ k \geqslant 0\ . \tag{7}$$

In Equation 7, non-negative k is called a ridge parameter or a regularization constant. Hoerl and Kennard have shown that the mean squared error for the ridge regression estimate $\hat{\boldsymbol{w}}^*$ is smaller than for $\hat{\boldsymbol{w}}$ [24]. The ridge regression estimate $\hat{\boldsymbol{w}}^*$ is also shorter than the maximum likelihood estimate $\hat{\boldsymbol{w}}$: $(\hat{\boldsymbol{w}}^*)^T(\hat{\boldsymbol{w}}^*) < (\hat{\boldsymbol{w}})^T(\hat{\boldsymbol{w}})$. This implies that the elements in $\hat{\boldsymbol{w}}^*$ are drawn towards zero. The mean squared error of the regression weight $\hat{\boldsymbol{w}}^*$ is

$$\begin{aligned}
\mathbb{E}[L_1^2(k)] &= \mathbb{E}[(\hat{\boldsymbol{w}}^* - \boldsymbol{w})^T(\hat{\boldsymbol{w}}^* - \boldsymbol{w})] && (8)\\
&= \mathbb{E}[(\boldsymbol{w} - \mathbb{E}[\boldsymbol{w}])^2] + (\mathbb{E}[\boldsymbol{w} - \hat{\boldsymbol{w}}^*])^2\\
&= \mathrm{Var}(\hat{\boldsymbol{w}}^*) + \mathrm{Bias}(\hat{\boldsymbol{w}}^*, \boldsymbol{w})^2\\
&= \beta^{-1}\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j + k)^2} + k^2\boldsymbol{w}^*(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I})^{-2}\boldsymbol{w}^*\ . && (9)
\end{aligned}$$

In Equation 8, λ_j is the jth eigenvalue of the matrix $\boldsymbol{X}^T\boldsymbol{X}$. As can be seen from this equation, the mean squared error of the regression weight can be decomposed to two terms, the first being the total variance of the regression weight, $\mathbb{E}[(\boldsymbol{w} - \mathbb{E}[\boldsymbol{w}])^2]$, and the second being the square of the bias, $(\mathbb{E}[\boldsymbol{w} - \hat{\boldsymbol{w}}^*])^2$, which tells the deviation of the estimate from the value of true regression weight. The bias term will be zero when $k = 0$, leading to the same mean squared error as for the non-regularized ML estimate $\hat{\boldsymbol{w}}$. Therefore, the maximum likelihood estimate is called an unbiased estimate. As was shown above, the variance of the regression weight can be high in data having large p, small n setting and/or multicollinear variables. Ridge regression regularization reduces the total variance of the regression weights because k appears in the denominator of summed terms in the variance part of Equation 9. This is done at the cost of introducing a small bias.

In ridge regression, the optimization function L when minimizing the squared distance between the true and predicted target variables is

$$L = \sum_{i=1}^{N}(y_i - \sum_{j}^{p} w_j x_{ij})^2 + k\sum_{j}^{p} w_j^2\ . \tag{10}$$

The ridge regression estimate for the regression weights is obtained by minimizing the penalized mean squared error between observed and predicted values presented in Equation 10. In Equation 10, k is the regularization parameter. The rigde regression, also called 2-norm regularization, ℓ_2-penalization, or weight decay, discourages the weights to become large by shrinking them towards zero. It can be though as introducing a bound prior for w weights. As a result, parameters are less uncertain than the ordinary least squares estimates. Moreover, in ridge regression regularization, highly correlated variables get similar weights resulting in grouping of the variables. This is reasonable, for example, in the analysis of gene expression data which consists of co-expressed and co-regulated genes. However, ridge regression

regularization does not set some weights completely to zero, which would be desirable in certain situations. Pushing some weights completely to zero would improve the interpretability of the results because it would remove the irrelevant dimensions from the analysis. This would lead to a sparse solution. In ridge regression, one still has to analyse all the weights in the end.

In regularized statistical modelling, one has to find the optimal regularization parameter k. The regularization parameter can be found using cross-validation procedure [254]. In m-fold cross-validation, the data are divided into m parts of about equal size. Then one of these sets is used on its turn as model validation data and the rest m-1 of these sets are used as training data. The parameter k could be defined in some interval that includes all the possible values for the regularization parameter. Then for each of the parameter values, a model is learned using the training data, and validated using the validation data. The objective function to find the optimal regularization parameter could be, for example, the likelihood function of the validation data, given the parameters learned using the training data. The learning and validation with different sets is performed m times. The final value of the object function is the mean of the likelihood function values obtained for m validation data sets. Then the regularization parameter giving the best value of the object function is chosen. The number of divisions m is usually chosen to be 5 or 10 leading to 5-fold or 10-fold cross-validation.

3.3 Bayesian data analysis

Until now the classical or frequentist probabilistic data analysis has been discussed in this thesis. The frequentist data analysis considers the model parameters as fixed quantities, and estimates them by maximizing the likelihood function. In contrast, this chapter turns to Bayesian data analysis in which in addition to the observed random variables, the model parameters are also considered as random variables, and the uncertainty of them is represented by probability density functions. In frequentist data analysis, probability is defined as a ratio of favourable events to all trials as the number of trials approaches infinity. However, there are events that cannot be repeated infinitely many times or even few times, resulting in difficulties in estimating the probabilities. In Bayesian data analysis, the probability is considered as a measure of uncertainty, or as a degree of belief which can be quantified using numerical values, and manipulated by different rules. An example of a difference between frequentist and Bayesian data analysis is the treatment of error bars or confidence intervals that define the uncertainty of estimated model parameters. In frequentist framework, the value of a model parameter is estimated by an estimator, and the confidence interval on this estimate is obtained by considering the distribution of all possible data sets; the confidence interval implies that when repeating the experiment and gaining a new equal sized data set, the value of a parameter estimated from this new data will be within the limits of the confidence interval in $(1 - a) \times 100$ % of the repetitions where a is a pre-specified significance level. Conversely, in the Bayesian framework where the uncertainty of a model parameter is measured as a probability density function, the range where the parameter value

is with the probability of $(1 - a) \times 100$ %, can be directly determined from the density function. This range is called Bayesian credible interval, and it gives the uncertainty of a model parameter a clear definition.

Bayesian data analysis fits well for data sets having small number of samples, and when there is a risk of overfitting because the Bayesian treatment has an intrinsic regularization property. In other words, the models learned by Bayesian inference adapt to the correct model complexity during the learning. Therefore, the Bayesian models are resistant to overfitting. This is explained more detailed in Chapter 3.3.3. Other advantages of the Bayesian framework are easy handling of missing data, and natural ways of model comparison.

3.3.1 Bayes' theorem

This chapter presents the components of Bayesian data analysis. The joint probability distribution of two random variables α, and β, namely $p(\alpha, \beta)$, can be presented in two different ways using the product rule of probability, and the symmetry property of the joint distribution: $p(\alpha, \beta) = p(\beta, \alpha)$:

$$p(\alpha, \beta) = p(\alpha|\beta)p(\beta) = p(\beta|\alpha)p(\alpha) = p(\beta, \alpha) \ . \tag{11}$$

From Equation 11 the Bayes' theorem can be derived as

$$p(\alpha|\beta) = \frac{p(\beta|\alpha)p(\alpha)}{p(\beta)} \ . \tag{12}$$

The Equation 12 implies that the conditional distribution of α given β can be obtained when the conditional distribution of β given α, $p(\beta|\alpha)$, and the marginal distributions $p(\alpha)$ and $p(\beta)$, are known. Keeping in mind the data modelling task, α can be considered as the model parameters and β the observed data. Therefore, $p(\beta|\alpha)$ is recognized as the likelihood function, and $p(\beta)$ is obtained by integrating the product of $p(\beta|\alpha)$ and $p(\alpha)$ over all model parameters. The distribution $p(\alpha)$ is the probability distribution of the model parameters, and $p(\alpha|\beta)$ can be considered as the probability distribution of the model parameters after the data β is observed. The distribution $p(\alpha)$ is called a prior distribution over the model parameters, and it quantifies the uncertainty of the parameters before seeing the data. In contrast, the uncertainty of the model parameters after seeing the data, $p(\alpha|\beta)$, is called a posterior distribution of the model parameters. The normalization constant $p(\beta)$ guarantees that the posterior distribution is a proper probability distribution, i.e. integrates into one. As a result, the model parameters have a probability distribution that can be used to assess the parameter uncertainty. Bayes' theorem can be also seen as an update rule for the probability distribution of the parameter values after observing a single data point β. The posterior distribution can then be used as a prior for the next sample, and so forth.

The prior distribution makes an assumption of the model parameter values before observing the data. It determines the possible parameter values and their probabilities; for example, if a prior assigns zero value to some parameters, then the posterior of those parameters will also be zero. One the one hand, prior can be subjective: for

example, it can be given by an expert of the phenomenon under study; or it can be based on a common sense. On the other hand, prior can be uninformative indicating that it has only some influence on the posterior. When considering the Bayes' theorem as an update rule for the parameter uncertainty, the influence of prior has a strong effect when there are only few data points. This prevents the model from overfitting to the data. However, when obtaining more data, the likelihood starts to dominate over the prior. Therefore, choosing a prior reflecting wrong assumptions of the data do not bias the final results when there is much data available. However, a strong prior may bias the results when the number of samples is limited.

A posterior distribution is obtained from a prior by gaining new evidence from the data. The posterior distribution of the parameters can be investigated as such, or single parameters estimates can be obtained, for example, by picking the most probable parameter values. The most probable parameter values are called Maximum A Posteriori (MAP) estimates. In contrast to estimating single parameter values and a single model, and considering that as the right model, the posterior gives the probability distribution over several parameter values or models. Instead of considering the most probable model, the probability distributions of all of the possible models can be investigated, which avoids picking a single solution. For example, in predictive analysis, the probability of a new data point given the model parameters can be obtained by marginalizing the likelihood function over the model parameters, thus avoiding the overfitting issue originating from picking a single, most likely overfitted model. The variance of the predictive distribution is the sum of noise in the data, and the uncertainty associated with the model parameters.

3.3.2 Conjugate prior distributions

Priors can be informative or uninformative. When an uninformative prior is desirable, a uniform prior is chosen. The uniform prior assigns an equal probability for all parameter values. Multiplying the likelihood function with the uniform prior does not affect the posterior distribution as the introduction of a uniform prior is equal to the multiplication of the likelihood function by a constant, and new normalization coefficient can be calculated for the posterior. Therefore, using uninformative priors leads the MAP estimates to equal the maximum likelihood estimates.

In this thesis, informative or very vague priors are used. Multiplying the likelihood function with any randomly selected prior distribution might lead the posterior distribution to be intractable. In that case, the normalization constants are impossible to derive. Therefore, conjugate priors are usually used to simplify calculations. For a given likelihood, a prior can be sought that is conjugate to the likelihood function, so that the posterior distribution has the same form as the prior distribution. In this case the posterior distribution can be easily derived.

If the likelihood function is Bernoulli distribution, then the conjugate prior is the beta distribution. If the likelihood is univariate Gaussian, the mean μ of this distribution has Gaussian conjugate distribution and the variance parameter σ^2 has inverse-Gamma as its conjugate distribution. As a result, the posterior distributions of μ and σ^2 also follow the Gaussian and inverse-Gamma distributions, respectively.

If the likelihood is multivariate Gaussian, then the conjugate prior of the mean parameters is also multivariate Gaussian, whereas the conjugate distribution of the covariance matrix is inverse-Wishart distribution. The prior distributions for the model parameters also have parameters that define the prior distribution. For example, the Gaussian prior for μ can be parameterized by mean m_0 and variance s_0^2. [260] The values for these hyperparameters can be chosen to reflect the prior assumptions or knowledge before seeing the data, or they can be learned from the data by Empirical Bayes strategy [254], or a prior distribution can be set for them as well, leading to hyperpriors and hierarchical Bayesian models.

The usage of different priors leads to different assumptions and regularization effects. These issues are considered in the following chapter.

3.3.3 Model selection and regularization in Bayesian data analysis

Bayesian data analysis provides inherent regularization without the need to optimize the regularization parameters, for example, by computationally tedious cross-validation procedures. This can be demonstrated by considering the Bayesian treatment for linear regression model presented in Equation 2. In this case the number of features in the data set, or the number of elements in the regression vector $(p+1)$, is the measure of the total model complexity. The target variable y is a function of $\boldsymbol{x}$ and $\boldsymbol{w}$, with additive Gaussian noise ε as shown in Equation 13. The noise ε is assumed to be normally distributed with zero mean and a precision β.

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{x}\boldsymbol{w} + \epsilon = \sum_{j=0}^{p} x_j w_j + \epsilon \ . \tag{13}$$

Now y follows a Gaussian distribution with mean $\boldsymbol{x}\boldsymbol{w}$ and precision β. The likelihood function for the whole data set is

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{n} (y_i|\boldsymbol{x}_i\boldsymbol{w}, \beta^{-1}) \ . \tag{14}$$

A conjugate prior distribution for the weight vector $\boldsymbol{w}$ is given by a Gaussian distribution

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \ . \tag{15}$$

The mean of the prior distribution is chosen to be zero and the covariance matrix is isotropic with a single precision parameter α.

Using Gaussian conjugate prior for the weight vector $\boldsymbol{w}$, its posterior distribution is also a Gaussian, $N(\boldsymbol{w}|\boldsymbol{m_N}, \boldsymbol{S}_N)$ where the mean parameter $\boldsymbol{m}_N$ and the covariance matrix $\boldsymbol{S}_N$ are

$$\begin{aligned} \boldsymbol{m}_N &= \beta \boldsymbol{S}_N \boldsymbol{X}^T \boldsymbol{t} \\ \boldsymbol{S}_N^{-1} &= \alpha \boldsymbol{I} + \beta \boldsymbol{X}^T \boldsymbol{X} \ . \end{aligned} \tag{16}$$

The logarithm of the posterior distribution is

$$\ln p(\boldsymbol{w}|\boldsymbol{y}) = -\frac{\beta}{2} \sum_{i=1}^{n} \{y_i - \boldsymbol{x}_i\boldsymbol{w}\}^2 - \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} + \text{const.} \tag{17}$$

It can be shown that the Equation 17 equals the Equation 10. Therefore, the maximization of the posterior distribution with respect to $\boldsymbol{w}$ is equivalent to the minimization of the sum-of-squares error function, or the negative log likelihood with the addition of a quadratic regularization term with the regularization parameter $k = \frac{\alpha}{\beta}$ which is the ratio between the precisions of the prior and the likelihood. In other words, the MAP estimate of the weight vector equals the ML estimate of the penalized likelihood. As a conclusion, the Gaussian prior of the weight vector $\boldsymbol{w}$ can be seen as an additional regularization term which leads to the ridge regularization. Maximizing the posterior distribution moves the solution towards the prior, preventing the model from overfitting.

Usually it is sufficient to set small values for the prior parameters leading to a vague prior distribution, and get more accurate estimate during the Bayesian learning. A proper value of α can also be learned from the data before the Bayesian treatment, for example, by Empirical Bayes strategy [254]. However, this can be avoided by setting a prior distribution also for the precision parameter α. The prior distribution of α is called a hyperprior. Conjugate hyperprior for a precision parameter is Gamma distribution with hyperparameters α_0 and β_0. These parameters can be set to make the hyperprior as uninformative as possible. The distribution of the precision parameter α, as well as the distribution of the regularization parameter $\frac{\alpha}{\beta}$ adapt to their optimal values and the correct model complexity during learning. Therefore, in the Bayesian framework, the model complexity is learned using all available data in contrast to the classical regularization methods where the data were divided to training and validation data sets. Moreover, the exact posterior distribution of the precision parameter α, and the distribution of the regularization parameter k are usually not interesting, so these parameters can be marginalized out of the posterior distribution, i.e. the posterior can be averaged over them. The averaging over uninterested parameters is an important property of Bayesian data analysis.

The different priors, for example, for regularization in the Bayesian treatment are studied heavily nowadays but are not further considered in this thesis. The regularization in the Bayesian framework can be further investigated by considering future data prediction. A prediction of future data instance y given observed $\boldsymbol{x}$ for the linear Bayesian model is done by marginalizing over the posterior distribution of the weight vector:

$$p(y|\boldsymbol{y}, \alpha, \beta) = \int p(y|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\boldsymbol{y}, \alpha, \beta) \mathrm{d}\boldsymbol{w} \ . \tag{18}$$

The result of this integration implies that the future observations follow a Gaussian distribution

$$p(y|\boldsymbol{x}, \boldsymbol{y}, \alpha, \beta) = N(y|\boldsymbol{m}_N^T \boldsymbol{x}, \sigma_N^2(\boldsymbol{x})) \ . \tag{19}$$

The mean the Gaussian distribution in Equation 19 is $\boldsymbol{m}_N^T \boldsymbol{x}$, and the variance $\sigma_N^2(\boldsymbol{x})$ is

$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{x}^T \boldsymbol{S}_N \boldsymbol{x} \ . \tag{20}$$

The variables $\boldsymbol{m}_N$ and $\boldsymbol{S}_N$ are the posterior mean and variance of the weight vector $\boldsymbol{w}$, respectively, as presented in Equation 16. The first term in Equation 20 is the data variance, whereas the second term reflects the uncertainty associated with the parameters $\boldsymbol{w}$. This is a way of the Bayesian data analysis to manage the uncertainty associated with the modelling process itself. The derivation of the predictive distribution in Equation 18 can be seen as a weighted averaging over the multiple solutions, each having a high variance. This marginalization avoids the overfitting because the prediction is obtained by averaging over the whole posterior instead of using a single, most likely overfitted model. For more details, see [254].

Bayesian data analysis provides a rigorous formulation to the automatic model complexity control, and is thus a proper way to address large p, small n -problem. The regularization effect in the Bayesian framework emerges when the prior distribution prefers parameter values leading to simpler models, or to more likely models before seeing the data. The model complexity measured as an effective number of the parameters, or degrees of freedom, adapts automatically to the size of the data set. Moreover, the prior distribution can constrain the solution space, for example, by making independence assumptions between the variables. In the previous example of Bayesian ridge regression, isotropic Gaussian prior was assumed for the regression weight parameters. This indicates that the different elements of the weight vectors are assumed independent of each other and to have the same variance. In practise it is sometimes difficult to choose a proper prior; the prior needs to be flexible enough and enable to constrain the complexity according to the data.

Chapters 3.2 and 3.3 presented regularization applied to linear regression, an supervised statistical method. However, the work done in this thesis considers exploratory data analysis. In exploratory data analysis, there is no need to build a model for predicting future values, although the model's capability to generalise to future observations is still desirable. In exploratory data analysis, it is more important to study learned joint distribution of the model parameters, and investigate the relationships between observed features. Moreover, in exploratory data analysis, no additional data sets are usually provided which could be used to validate the learned model. Both classical frequentist and Bayesian regularization methods can still be applied also to unsupervised data analysis methods which are the topics of Chapters 3.4, 3.5 and 4.

3.3.4 Bayesian inference using variational approximation

For many models of practical interest, it is infeasible to evaluate the posterior distribution, or to compute expectations with respect to this distribution. This could be because of the high dimensionality of the variable space, or because the posterior distribution has a highly complex form for which expectations are not analytically tractable. In the case of continuous variables, the required integrations may not have closed form analytical solutions while the dimensionality of the space and the complexity of the integrand may prohibit numerical integrations. The techniques used to infer the Bayesian probabilistic models when the posteriors cannot be solved analytically include Markov Chain Monte Carlo (MCMC) methods, and a variational

approximation scheme. The Bayesian model applied in this thesis is inferred by a variational approximation method which scales well to large applications. Variational approximation methods are typically used when the model has a large number of unknown parameters because in such situations MCMC methods often become computationally prohibitive. Variational approximation methods approximate the true posterior distribution using a simpler probability density function that is factorised in a particular way with respect to groups of variables, or has a specific parametric form. As such, variational approximation methods never generate exact results, which is their weakness.

Variational approximation methods consider optimization of a functional over a set of functions. A function is a mapping that takes a value of a variable as an input and returns the value of the function as an output, whereas a functional can be defined as a mapping which takes a function as an input, and returns the value of the functional as an output. Entropy presented in Equation 29 is an example of a functional; it takes a probability distribution $p(x)$ as an input and returns the entropy as an output. A functional derivative corresponds to the derivative of a function; the functional derivative expresses the change in value of the functional in response to infinitesimal changes in the input function. The optimal function is obtained by exploring all possible input functions to find the one that maximizes, or minimizes, the functional. For more details, see [254].

Consider a full Bayesian model in which all parameters are given a prior distribution. The set of all model parameters is denoted by $\boldsymbol{Z}$. The set of all observed variables is denoted by $\boldsymbol{X}$. The probabilistic model specifies the joint distribution $p(\boldsymbol{X}, \boldsymbol{Z})$, and the goal is to find an approximation for the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$, as well as for the model evidence or marginal data likelihood $p(\boldsymbol{X})$. For any choice of a distribution $q(\boldsymbol{Z})$, the log marginal likelihood can be decomposed as

$$\begin{aligned}
\ln p(\boldsymbol{X}) &= \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}) \mathrm{d}\boldsymbol{Z} \qquad (21)\\
&= \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}) q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}) q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z}\\
&= \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z}\\
&= \mathcal{L}(q) + KL_D(q||p),
\end{aligned}$$

where $\mathcal{L}(q)$ and $KL_D(q||p)$ are defined as follows:

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} \qquad (22)\\
KL_D(q||p) &= -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \mathrm{d}\boldsymbol{Z} \,. \qquad (23)
\end{aligned}$$

In Equations 22 and 23, both $\mathcal{L}(q)$ and $KL_D(q||p)$ are functionals of the distribution $q(\boldsymbol{Z})$. $KL_D(q||p)$ is noticed as a Kullback-Leibler divergence between the distributions p and q. The Kullback-Leibler divergence is considered in more detail in Chapter 3.4. In brief, it is a measure of dissimilarity between probability densities p and q. The Kullback-Leibler divergence is always non-negative, so $\ln p(\boldsymbol{X}) \geq \mathcal{L}(q)$.

Therefore, $\mathcal{L}(q)$ is a lower bound of the log marginal probability of the likelihood function. For more information, see [254].

If distribution $q(\boldsymbol{Z})$ corresponds to the approximate posterior distribution of the model parameters, it can be found by maximizing the $\mathcal{L}(q)$ with respect to the distribution $q(\boldsymbol{Z})$; this is equivalent to minimizing the KL divergence, or maximizing the expectation of the complete data log likelihood, i.e $\log(\boldsymbol{X}, \boldsymbol{Z})$. If any possible choice for $q(\boldsymbol{Z})$ is allowed, then the maximum of the lower bound occurs when the KL divergence vanishes, implying that $q(\boldsymbol{Z})$ equals the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$. However, as was noted above, the model is usually so complex that working with the true posterior distribution is intractable. Therefore, a restricted family of distributions $q(\boldsymbol{Z})$ is considered, and the member of this family is sought for which the lower bound is maximized. The goal is to restrict the distribution family sufficiently that they comprise only tractable distributions, while at the same time allowing the family to be sufficiently rich and flexible that it provides a good approximation to the true posterior distribution. The restriction is imposed purely to achieve tractability, and as rich a family of approximation distributions as possible should be used. There is no over-fitting associated with highly flexible distributions because using more flexible approximations simply allows to approach the true posterior distribution more closely. Variational methods lead to approximate solutions because the range of functionals over which the optimization is performed is restricted. For example, one could consider only quadratic functions, or functions composed of a linear combination of fixed basis functions in which only the coefficients in the linear combinations can vary. In the case of applications to probabilistic inference, the restriction may, for example, take the form of factorization assumptions [254, 261, 262].

To restrict the family of distributions $q(\boldsymbol{Z})$, factorised distributions can be used. The elements in $\boldsymbol{Z}$ are partitioned into M disjoint groups that are denoted $\boldsymbol{Z}_i$ where $i = 1, \ldots, M$. Then it is assumed that the distribution q factorises with respect to these groups as shown in Equation 24. This is called a mean field approximation [254].

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i) \ . \tag{24}$$

No other assumptions are made than the one presented in Equation 24. In particular, there is no restriction on the functional forms of the individual factors $q_i(\boldsymbol{Z}_i)$. Amongst all distributions $q(\boldsymbol{Z})$ having the form of Equation 24, the distribution for which the lower bound $\mathcal{L}(q)$ is largest, is sought. The optimization of $\mathcal{L}(q)$ with respect to all of the distributions $q_i(\boldsymbol{Z}_i)$ is done by optimizing with respect to each of the factors in turn. To achieve this, the above Equation 24 is substituted into the equation of the lower bound (Equation 22), and the dependence on one of the

factors $q_i(\boldsymbol{Z}_j)$ denoted simply by q_j is dissected out as shown below:

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \{\ln p(\boldsymbol{X}, \boldsymbol{Z}) - \sum_i \ln q_i\} \mathrm{d}\boldsymbol{Z} \\
&= \int \prod_i q_i \ln p(\boldsymbol{X}, \boldsymbol{Z}) \mathrm{d}\boldsymbol{Z} - \int \prod_i q_i \sum_i \ln q_i \mathrm{d}\boldsymbol{Z} \\
&= \int q_j \{ \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i \mathrm{d}\boldsymbol{Z}_i \} \mathrm{d}\boldsymbol{Z}_j - \int q_j \ln q_j \mathrm{d}\boldsymbol{Z}_j + \text{const} \\
&= \int q_j \ln \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) \mathrm{d}\boldsymbol{Z}_j - \int q_j \ln q_j \mathrm{d}\boldsymbol{Z}_j + \text{const} \ .
\end{aligned} \tag{25}$$

The distribution $\tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ appearing in Equation 25 is defined as

$$\ln \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const} = \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i \mathbf{d}\boldsymbol{Z}_i + \text{const} \ . \tag{26}$$

The notation $\mathbb{E}_{i \neq j}[.]$ in Equation 26 denotes an expectation with respect to the q distributions over all variables $\boldsymbol{z}_i$ for $i \neq j$. Now suppose $q_{i \neq j}$ is kept fixed, and $\mathcal{L}(q)$ is maximized with respect to all possible forms of the distribution $q_j(\boldsymbol{Z}_j)$. This is easily done by recognizing that $\mathcal{L}(q)$ is the negative KL divergence between $q_j(\boldsymbol{Z}_j)$ and $\tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$. Thus maximizing the lower bound is equivalent to minimizing the KL divergence. A general expression for the optimal solution $q_j^*(\boldsymbol{Z}_j)$ is

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{const} \ . \tag{27}$$

The form of solution presented in Equation 27 provides the basis for applications of variational methods. It implies that the log of the optimal solution for factor q_j is obtained simply by considering the log of the joint distribution over the data and model parameters, and then taking the expectation with respect to all of the other factors q_i for $i \neq j$. The additive constant in 27 is set by normalizing the distribution $q_j^*(\boldsymbol{Z}_j)$. Thus if an exponential is taken on both sides and the distribution normalized

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})]) d\boldsymbol{Z}_j} \ . \tag{28}$$

In practise, the form in Equation 27 is more convenient to work with, and so normalization constant is reinstated where required. The set of Equations for $j = 1, \ldots, M$ represent a set of consistency conditions for the maximum of the lower bound subject to the factorization constraint. However, they do not represent an explicit solution because the expression on the right-hand side for the optimum $q_j^*(\boldsymbol{Z}_j)$ depends on expectations computed with respect to the other factors $q_i^*(\boldsymbol{Z}_i)$ for $i \neq j$. A consistent solution is found by first initializing all of the factors $q_i^*(\boldsymbol{Z}_i)$ appropriately and then cycling through the factors and replacing each in turn with a revised estimate given the current estimates for all of the other factors. Convergence is guaranteed because the lower bound is convex with respect to each of the factors $q_i^*(\boldsymbol{Z}_i)$ [254, 263].

There are several sources of inaccuracy when applying factorised approximation methods. First, there might be interesting dependencies between model parameters which are, however, lost in the factorization of the parameters to independent sets. For example, when applying variational approximation to multivariate Gaussian data, and the posterior of the features are assumed to be independent of each other, the mean can be correctly captured but the variance of the factorised distributions are controlled by the direction of smallest variance of the true parameters, and the other variances are significantly under-estimated. In addition, the covariance matrix of the data will be isotropic. Second, factorised variational approximation tends to give approximations to the posterior distribution that are too compact. Third, if there are multiple modes in the true posterior distribution, maximization of the lower bound ends up to one of the local optima. For more discussion, see [254].

3.4 Modelling dependencies

Until now, Chapter 3 has considered only analysis of one set of random variables. From now on, the fusion of several sets of variables is discussed. Modelling dependencies between several data sets is one way to formulate a data fusion approach. It is an unsupervised, exploratory method for data integration. Often the dependency between different data sets is considered to be the interesting and relevant part in the data sets. In this chapter, the term dependency is defined and its properties examined.

Statistical dependency is an antonym to a more familiar term, statistical independency. Independency can be defined for random variables. Two random variables x and y are statistically independent if and only if their joint probability distribution $p(x, y)$ is the product of the marginal probability distribution of the variables: $p(x, y) = p(x)p(y)$ for all x and y. Thus, the joint density factorises into a product of the marginal terms if the variables are independent. Independency is essentially binary; either two variables are independent or they are not. Dependency refers to a deviation from independency. It is a continuous quantity, and in addition to knowing that some variables are dependent, the strength of the dependency can be assessed.

Determining dependency requires understanding of the concepts of information theory. Function that represents the information content of a continuous random variable x is denoted as entropy $H(x)$

$$H(X) = -\int p(x) \log p(x) \mathrm{d}x \ . \tag{29}$$

Entropy is dependent of the probability density function of x, $p(x)$. It is the average amount of uncertainty or randomness that an unobserved random variable has. Entropy tells how much information is received on average when a random variable is observed. For more details, see [264].

An average additional information required to specify the value of x using density function $q(x)$, instead of the true density function $p(x)$, can be measured by

Kullback-Leibler divergence $\mathrm{KL_D}(p\|q)$

$$\mathrm{KL_D}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x \ . \tag{30}$$

The Kullback-Leibler divergence is a measure of dissimilarity between probability densities $q(x)$ and $p(x)$. It can be interpreted as the average inefficiency of assuming that the distribution of x is $q(x)$ when the true distribution in reality is $p(x)$ [264].

The Kullback-Leibler (KL) divergence between the joint density function $p(x, y)$ and the product of the marginal density functions $p(x)p(y)$ can be used to measure the dependency between the two random variables x and y. In this case, the KL-divergence measures the deviance of the joint distribution from the product of the marginal distributions; hence it measures the deviance from the independency. This measure is called mutual information $I(X, Y)$ [264]

$$I(X, Y) = \int_x \int_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y \ . \tag{31}$$

The mutual information measures the information shared by random variables X and Y. It is a measure describing how much knowing one of the variables reduces the uncertainty about the other. The mutual information can be seen as a decrease in the entropy of one variable when the value of the other variable is known. If X and Y are independent of each other the mutual information is zero. If the variables are identical, the mutual information equals the entropy of X or Y. In that case, if the value of X is known, the entropy of Y reduces to zero.

The mutual information is regarded as the standard definition for the strength of the statistical dependency. However, mutual information needs the joint distribution of the random variables to obtain an exact and accurate characterization of the dependency. In practise, only the data sets, X and Y are available. Therefore, a dependency measure defined for the data sets is needed. A classical measure of association or dependency between two univariate variables x and y is the Pearson's correlation [265, 266] defined as

$$\rho_{xy} = \frac{\mathbb{E}[(x - [x])(y - \mathbb{E}[y])]}{\sqrt{(\mathbb{E}([x - \mathbb{E}[x]])^2)}\sqrt{(\mathbb{E}([y - \mathbb{E}[y]])^2)}} \ . \tag{32}$$

Pearson correlation ρ_{xy} measures the linear dependency between random variables x and y. In Equation 32, $\mathbb{E}[.]$ denotes the expectation over the joint probability distribution $p(x, y)$. In practise, the expectations are often replaced by population means when estimating the correlation, giving the sample correlation coefficient r_{xy}

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \ . \tag{33}$$

In Equation 33, $\bar{x}$ and $\bar{y}$ are the sample mean of X and Y, respectively, n is sample size, and s_x and s_y are the sample standard deviations of X and Y, respectively. Values of correlation range from -1 to 1. The sign measures the nature, and the

absolute value measures the strength of the dependency. Correlation is zero for statistically independent variables. However, the converse is not generally true. For the special case of multivariate normal distribution, a zero correlation also implies independence.

Mutual information has a strong connection with multivariate normal distribution. It can be shown that for jointly normally distributed random variables x and y, the correlation and mutual information are related [267]: the mutual information between x and y can be presented as

$$I(X,Y) = -\frac{1}{2}\log(1-\rho_{xy}^2) \ . \tag{34}$$

In Equation 34, ρ_{xy} is the Pearson correlation between the variables. This implies that for jointly normal variables we can use correlation as a dependency measure without loss of generality.

One way to perform a data fusion task is to combine the data sources so that the variation in common between them is emphasized. The relevant and interesting thing is often the information shared by two or more data sources. The shared information in this thesis refers to a shared variation in the data sets or statistical dependencies. More accurate results might be obtained when concentrating only on information present in all of the data sets and ignore variation specific to individual data sets. This enables to find information that could only accidentally be revealed by looking at any of the single sources only. One might also be interested in the variation remaining in data sets after extracting the dependencies. In the later chapters, generative models are presented that have representation for both the dependencies and the data set-specific variation; the residual variation in the data sets after extracting dependencies can therefore be easily obtained.

The models that are used in this work are based on maximizing the dependency between the two data sets. Modelling the data by seeking linear projective transformations of the data sets, such that the correlations between the lower dimensional representations of the data are maximized leads to canonical correlation analysis which is a traditional method to fuse several data sets, and to perform dependency modelling.

3.5 Canonical correlation analysis

3.5.1 Classical canonical correlation analysis

Canonical correlation analysis (CCA) can be partly understood through principal component analysis (PCA) [254] because the vocabulary and computation of these two methods are similar. Both are also linear methods. Principal component analysis is a popular method for dimensionality reduction and visualization of the data. PCA is defined as an orthogonal projection of the data onto a lower-dimensional linear subspace denoted as a principal subspace, such that the variance of the projected data is maximized [268]. PCA seeks a single low dimensional representation of a single data set, whereas CCA considers two data sets and seeks a lower dimensional representation for both of them, so that the correlation between the separate

presentations is maximized. CCA is a traditional method to seek dependencies between two data sets. It can be used to determine whether two sets of variables are statistically independent of each other in a linear sense, or conversely, to determine the magnitude of the relationship between the two sets. CCA aims to find linear projections of the variables in the first set that are associated with the linear projections of the variables in the other set when the variables in both sets have been observed on the same experimental units. The association is found by maximizing the correlation between the linear projections of the two data sets.

CCA can be seen as the most general version of the traditional least-squares method for the analysis of data sets. The roles of the two sets of variables in CCA are symmetric; it is not known a priori if variation of the features in the first data set imply variation of the features in the other set, or vice versa. In other words, neither of the data sources is considered as a target variable set or as a predictor variable set. Nevertheless, using the results obtained by CCA, the values of variables in one data set can be predicted using the values of the variables in the other data set, and vice versa. When studying gene expression and metabolomics changes, symmetric analysis is reasonable; changes in gene expression cause changes in metabolic pathway activities, which ultimately affect the metabolite concentrations, but it is known that the metabolites also affect the expression of genes. As a result from CCA, we find out which variables from gene expression and metabolomics data are related to each other, and which variables are relevant in providing more insight into the biological experimental hypotheses. The nature of relationships between the two data sets is explained by measuring the relative contribution of each variable to the canonical relationships obtained.

The model family in CCA is linear projections, data are assumed normally distributed, and the dependency measure is Pearson's sample correlation coefficient. Let us consider again two sets of random variables X and Y and their corresponding data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. The data matrices are assumed standardized implying that the means of the features equal zero and the variances equal one. In the data matrices, rows correspond to variables and columns correspond to samples. The dimension of the variable X is p and the dimension of the variable Y is q. It is also assumed that $p \leq q$. The columns of the data matrices are paired and there are n samples. The sample covariance matrices for the variables X and Y are $\boldsymbol{S_{xx}}$ and $\boldsymbol{S_{yy}}$, respectively. The covariance matrix for the concatenated set of variables is

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S_{xx}} & \boldsymbol{S_{xy}} \\ \boldsymbol{S_{yx}} & \boldsymbol{S_{yy}} \end{bmatrix},$$

where $\boldsymbol{S_{xy}}$ and $\boldsymbol{S_{yx}}$ are the between-sets covariance matrices.

CCA is formulated by denoting the directions corresponding to the first dimension of the lower dimensional projections of the data sets X and Y as $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, respectively. These are called the first pair of canonical variates or canonical components. The task in canonical correlation analysis is to determine the canonical variates $\boldsymbol{u}_1 = \boldsymbol{a}_1^T \boldsymbol{X}$ and $\boldsymbol{v}_1 = \boldsymbol{b}_1^T \boldsymbol{Y}$ which are maximally correlated. The vectors $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ correspond to the first projection vectors for data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The dimension of the canonical variate vectors $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, is the sample

size n, so there is one element of a canonical variate corresponding to one sample. The maximization of the correlation between the canonical variate pair leads to an optimization problem

$$\begin{aligned} \text{maximize} \quad & \frac{\boldsymbol{a}_1^T \boldsymbol{S}_{xy} \boldsymbol{b}_1}{\sqrt{\boldsymbol{a}_1^T \boldsymbol{S}_{xx} \boldsymbol{a}_1} \sqrt{\boldsymbol{b}_1^T \boldsymbol{S}_{yy} \boldsymbol{b}_1}} \\ \text{subject to} \quad & \boldsymbol{a}_1^T \boldsymbol{S}_{xx} \boldsymbol{a}_1 = 1 \\ & \boldsymbol{b}_1^T \boldsymbol{S}_{yy} \boldsymbol{b}_1 = 1 \ , \end{aligned} \tag{35}$$

where the canonical correlation is maximized subject to the constraint that the length of the canonical variates equals one. In practice, one needs to maximize only the numerator of the canonical correlation, $\boldsymbol{a}_1^T \boldsymbol{S}_{xy} \boldsymbol{b}_1$. To maximize the correlation, a Lagrangian function L

$$L = \boldsymbol{a}_1^T \boldsymbol{S}_{xy} \boldsymbol{b}_1 - \frac{\lambda}{2}(\boldsymbol{a}_1^T \boldsymbol{S}_{xx} \boldsymbol{a}_1 - 1) - \frac{\mu}{2}(\boldsymbol{b}_1^T \boldsymbol{S}_{yy} \boldsymbol{b}_1 - 1) \tag{36}$$

is defined. The parameters λ and μ are Lagrangian multipliers.

Differentiation of the Equation 36 with respect to λ and μ gives $\lambda = \mu$ and an eigenvalue problem shown below:

$$\begin{cases} \boldsymbol{S}_{xx}^{-1} \boldsymbol{S}_{xy} \boldsymbol{S}_{yy}^{-1} \boldsymbol{S}_{yx} \boldsymbol{a}_1 = \lambda^2 \boldsymbol{a}_1 \\ \boldsymbol{S}_{yy}^{-1} \boldsymbol{S}_{yx} \boldsymbol{S}_{xx}^{-1} \boldsymbol{S}_{xy} \boldsymbol{b}_1 = \lambda^2 \boldsymbol{b}_1 \end{cases} .$$

The eigenvectors of the solution to this eigenvalue problem define the linear projection vectors $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ that form the first pair of canonical variates. The square root of the eigenvalue, i.e. λ, is the canonical correlation between $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$. The second canonical variate pair $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ is found by maximizing the canonical correlation between them with the unit length constraints for $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$, and orthogonality constraints between the subsequent canonical variates: $\boldsymbol{a}_1^T \boldsymbol{S}_{xx} \boldsymbol{a}_2 = 0$, $\boldsymbol{b}_1^T \boldsymbol{S}_{yy} \boldsymbol{b}_2 = 0$, and $\boldsymbol{a}_1^T \boldsymbol{S}_{xy} \boldsymbol{b}_2 = 0$. It can be shown that p pairs of canonical variates corresponding to the rows of matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ can be determined by solving the generalized eigenvalue problem

$$\begin{cases} \boldsymbol{S}_{xx}^{-1} \boldsymbol{S}_{xy} \boldsymbol{S}_{yy}^{-1} \boldsymbol{S}_{yx} \boldsymbol{A} = \Lambda^2 \boldsymbol{A} \\ \boldsymbol{S}_{yy}^{-1} \boldsymbol{S}_{yx} \boldsymbol{S}_{xx}^{-1} \boldsymbol{S}_{xy} \boldsymbol{B} = \Lambda^2 \boldsymbol{B} \end{cases} .$$

As a solution to the above eigenvalue problem, p pairs of canonical variates $\boldsymbol{U} = \boldsymbol{A}^T \boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{B}^T \boldsymbol{Y}$ are obtained having p rows corresponding to the dimensions of the canonical variates and n columns corresponding to samples. The matrix $\boldsymbol{A}$ is a p×p matrix and the matrix $\boldsymbol{B}$ is a q×p matrix in which the columns correspond to the projection vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$, respectively. The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are called projection matrices. Canonical variate pairs have the properties that the canonical variates $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ have the correlation λ_i, all $\boldsymbol{u}$s are uncorrelated with each other, all $\boldsymbol{v}$s are uncorrelated with each other and $\boldsymbol{u}_i$, and $\boldsymbol{v}_j$ are uncorrelated if i≠j. This orthogonality property of subsequent canonical components implies that different canonical variate pairs represent different relationships between the two sets of variables.

3.5.2 Significance of the canonical correlations

Canonical correlation R_i between the ith canonical variate pair is obtained as a square of the eigenvalues of the generalized eigenvalue problem. Canonical correlation measures the strength of the overall relationship between the corresponding canonical variate pair, and is always at least as large as the largest correlation between any pair of variables of the two sets. There are at most p canonical variate pairs and canonical correlations if $p = \min(p, q)$. The squared canonical correlation R^2 is an estimate of the amount of shared variance between the canonical variates. It expresses the proportion of variance in a canonical variate that is related to the other variate of the same pair. Squared canonical correlation cannot however be interpreted as the degree of relation between the sets of variables. Squared canonical correlation of 0.5 means that the pair of canonical variates share 50 % of their variance, not that the sets of variables share 50 % of their variance.

When there is a large number of variables in the data sets, and thus a large number of canonical correlations, the probability that a canonical correlation is nonzero just by change, is considerably high. Therefore, only those canonical variates are analyzed whose corresponding canonical correlation coefficients are statistically significant at some preselected level, for example, 0.05. The classical test to estimate significance of canonical correlation is Bartlett's test [269]. The test assumes that the sample population is multivariate normal, and the sample size is reasonably large. The requirement of large sample size renders the Bartlett's test unsuitable for the data analysis of this work, so an alternative significance test is used in this thesis. More suitable methods to assess the significance of a canonical correlation are permutation tests [270]. In a permutation test, the observed value of a test statistic, for example, a canonical correlation, is compared to the distribution of values obtained from a random sample. For CCA, a meaningful random sample is one that is normally distributed with the same variance as the original variables but with no dependencies between variables. In this work, a permutation test for CCA presented in the master's thesis of Tapio Rinnet is used to assess the significance of canonical correlations [271]. In this permutation test, the data is randomized in a Gaussian way. The test proceeds as follows:

1. Divide the data set $\boldsymbol{X}$ into $\boldsymbol{X}_{train}$ and $\boldsymbol{X}_{test}$, and similarly for $\boldsymbol{Y}$.

2. Perform CCA for $\boldsymbol{X}_{train}$ and $\boldsymbol{Y}_{train}$, and calculate the canonical correlation R_i between the canonical variates in the test data.

3. Create random samples $\boldsymbol{X}_{rand}$ from multivariate normal distribution with a diagonal covariance matrix in which the values in the diagonals equal to the columnwise variances of $\boldsymbol{X}_{train}$, and similarly for $\boldsymbol{Y}$.

4. Perform CCA for $\boldsymbol{X}_{rand}$ and $\boldsymbol{Y}_{rand}$, and calculate the correlation $R_{i,rand}$ there is between the canonical variates in the original test data.

5. Compare R_i to the distribution of $R_{i,rand}$. Those canonical variates $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$ are deemed significant where the R_i is significantly larger than $R_{i,rand}$ at the 5 % confidence level.

3.5.3 Limitations of the classical canonical correlation analysis

Canonical correlation analysis is sensitive to small changes in the data sets, and it overfits badly when the data dimensionality is close to the sample size, and/or when there are many correlating variables in the data sets. In that case, canonical correlation analysis detects artificially high canonical correlations. Overfitting is especially harmful in exploratory data analysis when the aim is to form hypotheses on data. When the data dimensionality is higher than the sample size, the traditional canonical correlation analysis cannot be performed at all as the data-specific sample covariance matrices become singular, and their inverses cannot be determined. The solutions to these problems are studied in Chapters 3.5.5 and 3.5.7. The probabilistic formulation of CCA presented in Chapter 3.5.7 as such does not solve the problems but makes possible the Bayesian treatment with all the necessary tools. One of the cons of canonical correlation analysis also is the difficulty to interpret the results which is discussed in the next chapter.

3.5.4 Interpretation of the results of canonical correlation analysis

The contents of the canonical variates are defined by the canonical projections or weights $\boldsymbol{A}$ and $\boldsymbol{B}$. They can be interpreted in a similar way as multivariate regression weights or factor score coefficients in factor analysis; canonical projections reflect the independent or unique contributions of each variable to each canonical variate. The values for the weights are unbounded, so their relative magnitudes for a particulate variate are informative. If the weights are reported for standardized data having zero mean and unit variance, the weights are comparable within a canonical component, as well as between different canonical variate pairs. However, if the variables in the data are multicollinear, the interpretation of the weights becomes difficult. The multicollinearity might be due to co-expressed and co-regulated genes in the gene expression data, for example. A variable may receive a small weight simply because it is highly correlated with another variable in the same set, even though both variables have high correlations with the variate. In this situation, is becomes less clear where the observed effects originate. Therefore, when the variables within a data set are correlated, the weights do not reveal the variable's variance accounted for the variates. For more details, see [272].

Instead of canonical weights, canonical loadings, i.e. canonical structure coefficients, can be investigated. Canonical loadings denote to the correlations of the original variables in either of the two data sets and either canonical variate in a given canonical variate pair. They are commonly used to interpret the results of canonical correlation analysis as they pertain to the overall correlation of the respective variables and canonical variates [273–275]. The canonical loadings are bounded from above by 1 and from below by -1. Moreover, they are standardized across the canonical variates and can be compared between variates. The values of canonical loadings drop off quickly for the later, less highly correlated variates. The canonical loadings provide information about the relative contributions of variables to each independent canonical relationship. The variables in each set contributing heavily to the shared variance between variables and canonical variates are said to be related

to each other. Some significance tests for weights and canonical loadings have been proposed [276] but they are not further considered in this thesis.

Canonical loadings can be calculated by considering the relations of the observed variables in one data set with the canonical variates in either set. When the correlations of observed variables in the first (second) set are calculated with the canonical variates of the first (second) set, the loadings are called intraset loadings. The squared intraset loadings give the proportions of each variable's variance that is accounted for a canonical variate in the same set. When the correlations of observed variables in the first (second) set are calculated with the canonical variates of the second (first) set, the loadings are called interset loadings. The squared interset loadings give the proportions of each variable's variance that is accounted for by a canonical variate of the other set. For more details, see [277].

In addition to investigating the significance of canonical correlations, it is useful to determine the amount of variation in the data sets explained by each canonical variate pair. It is possible that despite having found very strong and significant correlations between two data sets, they represent insignificant variation in the original data sets. Therefore, it is important to quantify how strongly a canonical variate interacts with its own data set. A useful measure for this is the squared intraset loading. Squared canonical intraset loading equals variable's variance shared with the corresponding canonical variate. For both canonical variates in a canonical component, an average proportion of the variance extracted by them can be calculated as a mean of canonical intraset loadings of individual variables. These can be further summed to have a single measure of the extent to which the canonical variates account for the variation in their respective sets. For more information, see [272, 278].

Another measure to assess the variance in the data sets accounted for canonical variate pairs is redundancy. Redundancy describes the actual variability in one set of variables explained by the other. Redundancy depicts how redundant one set of variables is given the other set of variables. In other words, redundancy can be used to determine the amount of variance shared by two sets of variables. Together with the average squared canonical intraset loadings, redundancies also reveal the data set that has more relevant (shared) information. Redundancy is analogous to multiple regression's R^2 statistics, the square of the correlation coefficient. Redundancy measure is obtained when the squared canonical correlation coefficient is multiplied by the variance extracted, i.e. the average of squared canonical intraset loadings over variables. Redundancies of canonical variates in relation to the corresponding data sets are presented in Equation 37. Redundancies can also be summed or averaged over the significant canonical variate pairs to obtain a single index of redundancy. Redundancy is also useful in assessing practical significance of canonical correlation coefficients. For more details, see [272, 279].

$$\mathrm{Rd}_{X,u_i} = \sum_{j=1}^{p} \frac{\mathrm{cor}^2(x_j, u_i)}{p} \times R_i^2 \tag{37}$$

$$\mathrm{Rd}_{Y,v_i} = \sum_{j=1}^{q} \frac{\mathrm{cor}^2(y_j, v_i)}{q} \times R_i^2$$

Graphical representations and visualization are crucial to help interpreting the results obtained by CCA. Scatter plots of the canonical variates in two dimensions reveal the separation of the samples to clusters based on their characteristics. The features responsible for the clustering of the samples can be revealed by investigating the canonical loadings of individual variables. The canonical loadings can be visualized using so called correlation circles where the coordinates of the variables are determined by the canonical loadings corresponding, for example, two given canonical variates. In a correlation circle, all variables end up inside a circle of radius 1. Correlation circles help revealing the correlations between two sets of variables, and indicate clusters of highly correlating variables. Furthermore, because the canonical variate vectors are orthogonal, there is a small degree of overlap between the variables contributing to the differences in samples modelled by different dimension of canonical variates. Thus each canonical variate pair focuses on a specific aspect of the data set. The canonical variates and correlation circles can be plotted either by using the canonical variates corresponding to the data set $\boldsymbol{X}$ or the data set $\boldsymbol{Y}$. If the canonical correlation is high for the corresponding canonical variates, the plots obtained using either set should be almost similar. In addition, a visualization tool called eye diagram is used in this thesis to represent the content of canonical components [280].

3.5.5 Regularized canonical correlation analysis

Classical canonical correlation analysis can be applied only to data having more samples than variables. In that case the data matrices are of full column rank, i.e. the rank of the data matrices equals the data set dimensions, p and q, respectively, and the sample covariance matrices are nonsingular. When the number of variables increases, greatest canonical correlations are nearly 1 due to overfitting. Therefore, a standard condition where CCA can be reliably perform is $n \geq p + q + 1$ [281]. Multicollinearity of the variables within the data sets also causes data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ not to be of full rank. As a result, the covariance matrices can be singular. Singular and ill-conditioned covariance matrices cannot be inverted, or their inverses are at least unstable, i.e. they have a high variance. To overcome these drawbacks of applying CCA to high dimensional and multicollinear data, regularization methods presented in the context of regression in Chapter 3.2 can be applied to canonical correlation analysis.

Ridge regression regularization for canonical correlation analysis was first introduced by Vinod [25] and further developed in [282]. In the ridge regularization framework for CCA, the sample covariance matrices $\boldsymbol{S}_{xx}$ and $\boldsymbol{S}_{yy}$ are replaced by

matrices $\boldsymbol{\Sigma}_{xx}(k_1)$ and $\boldsymbol{\Sigma}_{yy}(k_2)$, respectively, defined as

$$\boldsymbol{\Sigma}_{xx}(k_1) = \boldsymbol{S}_{xx} + k_1\boldsymbol{I}_p \text{ and } \boldsymbol{\Sigma}_{yy}(k_2) = \boldsymbol{S}_{yy} + k_2\boldsymbol{I}_q \ . \tag{38}$$

The values for regularization parameters k_1 and k_2 are chosen so that the generalization ability of the model is maximized. This can be done using cross-validation or bootstrapping -methods [283, 284]. In this thesis, cross-validation procedure is used. The possible values of the regularization parameters are defined in a two-dimensional grid, and for every combination of the regularization parameters, 10-fold cross-validation is performed to obtain the average value of the optimization function. The regularization parameters maximizing the objective function are finally chosen to be the regularization parameters of the final analysis.

In cross-validation, the objective function used to maximize the generalization performance of the validation data is usually the data likelihood. In addition to likelihood function, several other objective functions can be used for canonical correlation analysis, such as mutual information, or the sum of all canonical correlations, which are maximized [38]. In this thesis, the maximization of the mutual information of the validation data is used. Other object functions could also be considered such as minimizing the mean difference between the canonical correlation in the training and validation data [36, 285], or minimizing the mean squared prediction error (MSPE) of the canonical variates [286]. The performance of the different object functions is somewhat covered in [36, 38, 285, 286] but a comprehensive comparison of these methods in optimizing the regularization parameters in canonical correlation analysis has not been conducted. This issue is not further considered in this thesis.

Ridge regression shrinks the canonical weights by imposing a penalty on their size. The diagonal of the covariance matrices are penalized such that they become nonsingular and give unique estimates for the model parameters. To perform the regularized canonical correlation analysis properly, the regularization parameters k_1 and k_2 should not be considered equal for each row of the covariance matrix but unequal. However, in this case the optimization of the regularization parameters would become computationally very tedious. Practically, the regularization parameters can be considered equal for all rows. In regularized canonical correlation analysis developed and implemented by González et al. [26, 27, 287], an equal regularization parameter is considered, and it is optimized using cross-validation by maximizing only the first canonical correlation for the validation data. The method has been applied, for example, to analyse gene expression measures of 120 genes, and gas chromatography mass spectrometric measurements of 21 fatty acids in the liver cells of 40 mice [287–289]. The goal of this research was to investigate the effect of the genotypes (mutant and wild type), and five different oil diets on gene expression and fatty acid levels in liver cells. Regularized canonical correlation analysis was shown to yield the same results obtained by classical multivariate data analysis methods, such as hierarchical clustering and principal component analysis applied to the two data sets separately but, in addition, the regularized CCA revealed findings not visible in single data set analysis. In different experimental conditions, multiple metabolite levels were seen to correlate with the expression of genes coding enzymes

of the fatty acid metabolites and with the expression of some transcriptional factors. The authors in [287] hypothesized regulatory relationships between correlated metabolites and genes, and proposed a link between a certain fatty acid and the activity of a certain transcriptional factor pathway which will be target of the future validating experiments.

Jozefczuk et al. [290] have applied ridge regression regularized CCA to integrate metabolite data together with gene expression to provide an insight into system level stress adjustments of *E. Coli*. The authors obtained transcriptomics and metabolite responses to five different perturbations of the standard growing conditions (cold, heat, oxidative stress, lactose diauxie and stationary phase). The data was measured in several time points during the fermentation. The analysis consisted on 188 metabolites and 288 genes, whereas the total sample number was more than 550. Canonical correlation analysis was performed for the metabolite and transcriptomics data on each perturbation separately but the number of samples obtained from each perturbation experiment was not mentioned. Nevertheless, the analysis identified a number of significant condition-dependent associations between metabolites and transcripts; authors found groups of correlating metabolites, metabolite genes and transcriptional factors which are known to regulate or potentially regulate the metabolic processes. The co-occurrence of metabolic and transcript responses for functionally related genes and metabolites was proposed to be an effect of strong co-regulation on both levels of responses. These associations were found for previously known specific pathway regulations, as well as potential new ones that will be targets for future research. In conclusion, according to the authors, CCA is a successful explorative tool to display associations between genes and metabolites that are less prominent by means of direct linear relationships (e.g. Pearson correlation) in the initial data [287, 290].

Applying the classical ridge regression regularization technique to canonical correlation requires the optimization of the regularization parameters using cross-validation procedure. However, optimizing the regularization parameters can be unambiguous and computationally very tedious. In the studies of [26, 27, 287, 290] the number of features was not comparable to the number usually obtained from the high-throughput bioinformatics studies or the number of samples was very large. Therefore, it is questionable, whether the ridge regression regularized CCA generalizes well to large-scale data sets without overfitting.

3.5.6 Generative latent factor model for canonical correlation analysis

This section presents a visualization tool for probabilistic models called a plate diagram. The plate diagrams can be used to understand the probabilistic canonical correlation analysis presented in Chapter 3.5.7, and the models derived from that. A plate diagram is a visual representation of a joint probability distribution of a probabilistic generative model. It shows the form and properties of the distribution through a diagrammatic representation. This visualization eases especially the understanding of very complex probabilistic model structures. The structure of the graph is called generative because it reflects the process whereby the observed data

has been arisen. The generative graph is composed of nodes and edges connecting the nodes. A node corresponds to a random variable, and is associated with the corresponding conditional probability distribution. An edge connecting two nodes can have a direction denoted by an arrow, and if cycles formed by the edges are not allowed, the graph is called directed acyclic graph. A directional edge corresponds to a conditional dependency of the variables it connects; for example, if there is an edge from node A to node B, then the probability distribution of B is conditioned on the values of A. Moreover, A is called a parent node and B is a child node. For more details, see [254].

A generative model and its plate diagram can consist of three types of random variables: observed variables corresponding to the observed data, model parameters, and latent or hidden variables. Observed variables are shaded in the plate diagram, and if there are several observations of the same random variable, a plate (rectangle) is drawn around the corresponding node, and the sample size N is marked to the right bottom corner of the plate. In contrast, the model parameters are not observed but estimated from the data. The model parameters are usually of interest because their values depict the underlying process being modelled. Moreover, the latent variables are also not observed but are sometimes introduced to allow construction of complex joint distributions of the observed variables and model parameters from simpler components, i.e. the conditional distributions. The latent variables are not necessarily interesting, and they might have no interpretation. They can be also considered as generating the data through an unknown mapping determined by the model parameters. For more details, see [254].

Consider a plate diagram for a generative latent variable model shown in Figure 9 [31]. A plate is drawn around the graph implying that there are N instances of $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Generally, the generative graph determines the joint distribution of all random variables in the model, and the factorization of the joint distribution to a product of conditional distributions. The joint distribution $p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ of the model in Figure 9 factorises as follows:

$$p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = p(\boldsymbol{X})p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Y}|\boldsymbol{Z}) \ . \tag{39}$$

The joint distribution is a product of the conditional distributions associated with the graph nodes. It is notable that the variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent of each other given the value of the latent parameter $\boldsymbol{Z}$.

The edges in the plate diagram show the conditional dependence and independence properties of the random variables. Therefore, the edges can be considered to express the probabilistic and causal relationships between the random variables [254]. Particularly, the absence of edges between the nodes usually conveys interesting information about the properties of the distribution that the graph represents. Moreover, the number of model parameters can be reduced by dropping the edges in the graph; this leads to restriction of the possible distributions that the graph can encompass. An alternative way to reduce the number of independent model parameters is by sharing them. This is also denoted as tying the parameters. An example of tying of parameters is seen in Figure 9 where the conditional distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ share the latent variable $\boldsymbol{Z}$.

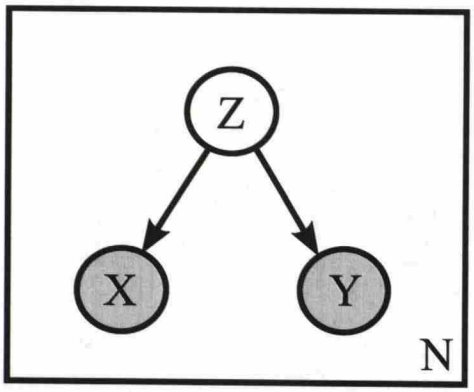


Figure 9: Probabilistic graphical model for canonical correlation analysis. [31]

Plate diagrams have various advantages and applications: they can be used to design and motivate new probabilistic models; they show the conditional independence properties of the variables; and their structure tells how synthetic data points can be sampled from the model. Sampling is started from the nodes with no parents by drawing random samples from their corresponding probability distributions. This process continues by sampling the values of other nodes from their conditional distribution given the earlier sampled parent values. This method is called ancestral sampling. For more discussion, see [254].

A plate diagram in Figure 9 is for a generative latent variable model corresponding to the probabilistic canonical correlation analysis, formulated in [29]. The nodes $\boldsymbol{X}$ and $\boldsymbol{Y}$ correspond to the two data sets and $\boldsymbol{Z}$ is a shared latent variable between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For each pair of $\boldsymbol{X}$ and $\boldsymbol{Y}$ there is a shared instance $\boldsymbol{Z}$.

3.5.7 Probabilistic canonical correlation analysis

In addition to the graph structure giving the conditional independence assumptions for the random variables, probability distributions for each node are needed to determine the full model. Following [29], the nodes in the graph presented in Figure 9 could have the following probability distributions:

$$\begin{aligned} \boldsymbol{z} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ \boldsymbol{x}|\boldsymbol{z} &\sim \mathcal{N}(\boldsymbol{W}_x\boldsymbol{z} + \boldsymbol{\mu}_x, \boldsymbol{\Psi}_x),\ \boldsymbol{W}_x \in \mathbb{R}^{p \times d_z} \\ \boldsymbol{y}|\boldsymbol{z} &\sim \mathcal{N}(\boldsymbol{W}_y\boldsymbol{z} + \boldsymbol{\mu}_y, \boldsymbol{\Psi}_y),\ \boldsymbol{W}_y \in \mathbb{R}^{q \times d_z}\ . \end{aligned} \tag{40}$$

In Equation 40, $\boldsymbol{z}$ follows a zero mean Gaussian distribution with identity covariance matrix. The dimensionality of $\boldsymbol{z}$, d_z, is equal to p, the minimum dimensionality of the two data sets. The mean of $\boldsymbol{x}$ consists of an overall mean $\boldsymbol{\mu}_x$, and a linear function of the shared variable $\boldsymbol{z}$ through a transformation matrix $\boldsymbol{W}_x$. The covariance matrix of $\boldsymbol{x}$ is $\boldsymbol{\Psi}_x$. Similar distributions can be set to the random variable $\boldsymbol{y}$ corresponding to the data set $\boldsymbol{Y}$.

In classical canonical correlations analysis, no probabilistic treatment of the learned parameters is usually performed. Nevertheless, Bach and Jordan [29] showed

that the maximum likelihood solution for the model having the structure presented in Figure 9 and the probability distributions presented in Equation 40 corresponds to the classical canonical correlation analysis. Theorem 2 in [29] states that the maximum likelihood estimates for the model parameters are

$$\begin{aligned} \hat{\boldsymbol{W}}_x &= \boldsymbol{\Sigma}_{xx}\boldsymbol{U}_x\boldsymbol{M}_x, \\ \hat{\boldsymbol{W}}_y &= \boldsymbol{\Sigma}_{yy}\boldsymbol{U}_y\boldsymbol{M}_y, \\ \hat{\boldsymbol{\Psi}}_x &= \boldsymbol{\Sigma}_{xx} - \hat{\boldsymbol{W}}_x\hat{\boldsymbol{W}}_x^T \\ \hat{\boldsymbol{\Psi}}_y &= \boldsymbol{\Sigma}_{yy} - \hat{\boldsymbol{W}}_y\hat{\boldsymbol{W}}_y^T . \end{aligned} \tag{41}$$

In Equation 41, $\boldsymbol{U}_x$ and $\boldsymbol{U}_y$ denote the CCA projection or weight matrices. The matrices $\boldsymbol{M}_x$ and $\boldsymbol{M}_y$ are arbitrary matrices with spectral norms smaller than one, such that $\boldsymbol{M}_x\boldsymbol{M}_y^T = \boldsymbol{P}$ where $\boldsymbol{P}$ is a diagonal matrix having the canonical correlation on its diagonal. The expectations of $\boldsymbol{z}$ given $\boldsymbol{x}$ or $\boldsymbol{y}$, $\mathbb{E}(\boldsymbol{z}|\boldsymbol{x})$ and $\mathbb{E}(\boldsymbol{z}|\boldsymbol{y})$, lie on the CCA projection space, i.e. the subspace that corresponds to the space spanned by the CCA projections [29]. However, the solution presented in 41 has a rotational ambiguity caused by $\boldsymbol{M}_x$ and $\boldsymbol{M}_y$ implying that the true canonical components corresponding to the classical solution are not revealed. [291] presents a method to solve this ambiguity and to find the actual CCA projections. The CCA projection matrices corresponding to the classical CCA results are

$$\begin{cases} \boldsymbol{U}_{xd} = \boldsymbol{\Sigma}_{xx}^{-1}\hat{\boldsymbol{W}}_x(\boldsymbol{I}_d - \boldsymbol{B}_x^{-1})^{-\frac{1}{2}}\boldsymbol{R}_x, \\ \boldsymbol{U}_{yd} = \boldsymbol{\Sigma}_{yy}^{-1}\hat{\boldsymbol{W}}_y(\boldsymbol{I}_d - \boldsymbol{B}_y^{-1})^{-\frac{1}{2}}\boldsymbol{R}_y. \end{cases}$$

Here $\boldsymbol{B}_x = \hat{\boldsymbol{W}}_x^T\hat{\boldsymbol{\Psi}}_x\hat{\boldsymbol{W}}_x + \boldsymbol{I}_d$ and $\boldsymbol{B}_y = \hat{\boldsymbol{W}}_y^T\hat{\boldsymbol{\Psi}}_y\hat{\boldsymbol{W}}_y + \boldsymbol{I}_d$. The matrix $\boldsymbol{R}_x$ contains the eigenvectors of $(\boldsymbol{I}_d - \boldsymbol{B}_x^{-1})^{\frac{1}{2}}(\boldsymbol{I}_d - \boldsymbol{B}_y^{-1})(\boldsymbol{I}_d - \boldsymbol{B}_x^{-1})^{\frac{1}{2}}$ and the matrix $\boldsymbol{R}_y$ contains the eigenvectors of $(\boldsymbol{I}_d - \boldsymbol{B}_y^{-1})^{\frac{1}{2}}(\boldsymbol{I}_d - \boldsymbol{B}_x^{-1})(\boldsymbol{I}_d - \boldsymbol{B}_y^{-1})^{\frac{1}{2}}$. For more details, see [291].

In [29], Bach and Jordan also derive the minimum of the negative log likelihood for the probabilistic canonical correlation analysis shown below:

$$\min\{-\log L\} = \frac{(p+q)n}{2}log\,2\pi e + \frac{n}{2}log|\hat{\boldsymbol{\Sigma}}_x x| + \frac{n}{2}log|\hat{\boldsymbol{\Sigma}}_y y| + \frac{n}{2}\sum_{i=1}^{d_z} log(1-\rho_i^2) \ . \tag{42}$$

The last term of the minimum of the negative likelihood is the negative mutual information. Thus, minimizing of the negative log likelihood with respect to the canonical correlation ρ_i corresponds to maximizing the mutual information between the two data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$. Hence using the mutual information as a optimization function when determining the regularization parameters for ridge regression CCA using cross-validation is reasonable.

Klami and Kaski [30, 31] have introduced another generative model that also corresponds to the classical canonical correlation analysis. The plate diagram for this model is shown in Figure 10. In addition to the shared latent variable $\boldsymbol{Z}$, the model includes also latent variables specific to the two data sets, $\boldsymbol{Z}_x$ and $\boldsymbol{Z}_y$. The

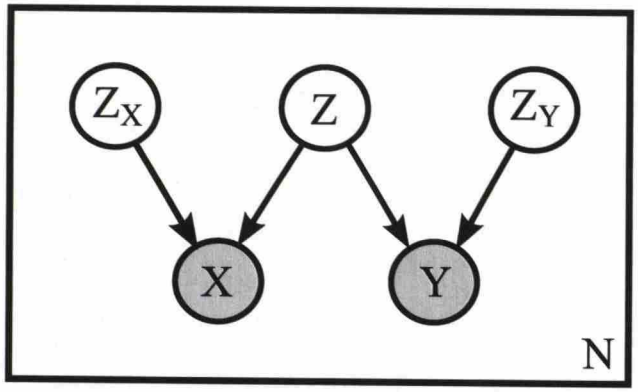


Figure 10: Probabilistic graphical model for canonical correlation analysis including data set-specific latent variables [30, 31].

conditional probability distributions for this model are

$$\begin{aligned} \boldsymbol{z}, \boldsymbol{z}_x, \boldsymbol{z}_y &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ \boldsymbol{x}|\boldsymbol{z}, \boldsymbol{z}_x &\sim \mathcal{N}(\boldsymbol{W}_x\boldsymbol{z} + \boldsymbol{B}_x\boldsymbol{z}_x + \boldsymbol{\mu}_x, \sigma_x^2\boldsymbol{I}),\ \boldsymbol{W}_x \in \mathbb{R}^{p\times d_z}\ \boldsymbol{B}_x \in \mathbb{R}^{p\times d_{z_x}} \\ \boldsymbol{y}|\boldsymbol{z}, \boldsymbol{z}_y &\sim \mathcal{N}(\boldsymbol{W}_y\boldsymbol{z} + \boldsymbol{B}_y\boldsymbol{z}_y + \boldsymbol{\mu}_y, \sigma_y^2\boldsymbol{I}),\ \boldsymbol{W}_y \in \mathbb{R}^{q\times d_z}\ \boldsymbol{B}_y \in \mathbb{R}^{p\times d_{z_y}}\ . \end{aligned} \tag{43}$$

Note that in Equation 43, $\boldsymbol{B}_x$ and $\boldsymbol{B}_y$ are different from the $\boldsymbol{B}_x$ and $\boldsymbol{B}_y$ shown in equations above.

The data set-specific latent variable $\boldsymbol{Z}_x$ (and corresponding $\boldsymbol{Z}_y$) accounts for the variation existing only in data set $\boldsymbol{X}$ ($\boldsymbol{Y}$). Variation in either data set can have its own characteristic ways of producing noise and irrelevant information to the corresponding data set. In a generative sense, for example, the data $\boldsymbol{X}$ is generated from the shared source of variation, and the data-specific variation as a linear function through the matrices $\boldsymbol{W}_x$ and $\boldsymbol{B}_x$. Moreover, the non-shared variation may be also of interest and it can be modelled using the above model. However, the shared variation is often considered more relevant, and the data-specific variation is ignored. The data set-specific latent variables can be considered as nuisance parameters, and their modelling is avoided by marginalizing them out of the model. When the data set-specific variation has a full dimensionality denoting that the dimensionality of the data set-specific latent variables match the corresponding data dimensionality, $d_{z_x}+1 = p$ and $d_{z_y}+1 = q$, and the data set-specific latent variables are marginalized out, the model in Figure 10 corresponds also to the classical canonical correlation analysis [30, 31]. Then the covariance matrix for data set $\boldsymbol{X}$, $\boldsymbol{\Psi}_x$, can be presented as $\boldsymbol{B}_x\boldsymbol{B}_x^T + \sigma_x^2\boldsymbol{I}$ which is a covariance matrix of rank $d_{z_x}+1$. The corresponding fact is true for the covariance matrix of $\boldsymbol{y}$. If the data set-specific latent variables are of full dimensionality, then the residual variances σ_x^2 and σ_y^2 are both zero. As a result this leads to the same model as proposed in the work of Bach and Jordan [29].

Several things support the marginalization of the data set-specific latent variables from the model: First, if they are included in the modelling, the computation would be inefficient comparing to the situation when they are integrated out. Second, the

marginalization is a crucial step to find the correct shared latent signal; the model including the data set-specific latent signals is very flexible, and there would be a serious risk of overfitting these less interesting signals. In that case, the data set-specific latent signals could prefer modelling also some of the shared signal and the dependencies would not be detected correctly. By introducing the data set-specific latent variable which correspond to the marginal densities of the data sets, and integrating them out allows to use likelihood more generally as a score function for fitting dependency seeking models. For more discussion, see [30, 31].

The data set-specific latent variables can be found as a pre-processing step after the model parameters $\hat{\boldsymbol{W}}_x$, $\hat{\boldsymbol{W}}_y$, $\hat{\boldsymbol{\Psi}}_x$ and $\hat{\boldsymbol{\Psi}}_x$ have been learned. The covariance matrix $\hat{\boldsymbol{\Psi}}_x$ can be factorised by learning the model

$$\boldsymbol{x}_j \sim \mathcal{N}(\boldsymbol{B}_x \boldsymbol{z}_{x_j}, \boldsymbol{\Sigma}_x + \sigma_x^2 \boldsymbol{I}) \ . \tag{44}$$

In Equation 44, $\boldsymbol{z}_{x_j}$ is the data set-specific latent variable and $\boldsymbol{\Sigma}_x = \hat{\boldsymbol{W}}_x^T \hat{\boldsymbol{W}}_x$ is the variation already explained by the shared latent variable. This model is essentially same as the probabilistic principal component analysis [292].

The probabilistic canonical correlation analysis has several advantages and applications: it deepens the understanding of CCA; it enables the use of local canonical correlation models as components of larger hierarchical probabilistic models; and it suggests generalization of CCA to members of the exponential family other than the Gaussian distribution [29, 30, 293]. However, the probabilistic canonical correlation analysis cannot be applied when the dimensionality of the variables is higher than the sample size, the same problem encountered by classical canonical correlation analysis. The overfitting in both cases is serious, or the method cannot be applied at all. The problem resides especially in the estimation of the large covariance matrices of the data sets having large p but small n. Nevertheless, the probabilistic framework enables the usage of Bayesian probabilistic modelling tools discussed in Chapter 3.3. The Bayesian treatment makes possible the use of priors which regularize the model, and applying Bayesian regularization can be more straightforward and efficient than the classical regularization methods. In addition, the probabilistic canonical correlation analysis could have less parameters to learn if a compromise relating to the dimensionality of the data set-specific latent variables (i.e. the number of columns in matrix $\boldsymbol{B}_x$) is made. This leads to the constraining of the data covariance matrices by decreasing the rank of the covariance matrix $\boldsymbol{\Psi}_x$. However, when the data set-specific latent variables are not of full dimensionality, they begin to also model the shared variation, and the shared latent variable begins to model the data set-specific variation as noted above. Nevertheless, in this case the model could still detect some of the shared variation right and perform better with data sets having small n and large p. A model that utilize is approach is discussed in a section 4.

3.5.8 Bayesian canonical correlation analysis

Probabilistic canonical correlation analysis presented in Chapter 3.5.7 can be readily extended to the Bayesian framework where prior distributions are defined for the

model parameters. Consider model in Figure 11, in which in addition to the observed and latent variables, the nodes corresponding to model parameters are added. Before introducing the prior distributions, a feature-wise concatenation of the data $\boldsymbol{V} = [\boldsymbol{X}; \boldsymbol{Y}]$ is performed, and the model is written based on this concatenation [294]. The CCA model can then be written

$$\begin{aligned} \boldsymbol{z} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ \boldsymbol{v}|\boldsymbol{z} &\sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Psi}),\ \boldsymbol{W} \in \mathbb{R}^{(p+q)\times d_z}\ . \end{aligned} \tag{45}$$

In Equation 45, $\boldsymbol{\mu}$ is the concatenation of $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$, $\boldsymbol{W}$ is the concatenation of the matrices $\boldsymbol{W}_x$ $\boldsymbol{W}_y$, $\boldsymbol{W} = [\boldsymbol{W}_x, \boldsymbol{W}_y]$, and $\boldsymbol{\Psi}$ is a block diagonal matrix having $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ on its diagonal.

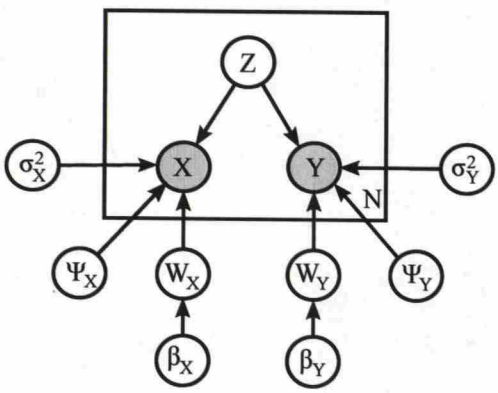


Figure 11: Probabilistic graphical model for Bayesian canonical correlation analysis.

The prior distribution for the parameters in the model 45 are

$$\begin{aligned} \boldsymbol{\mu} &\sim \mathcal{N}(0, \sigma^2\boldsymbol{I}), \\ \boldsymbol{\Psi}_x, \boldsymbol{\Psi}_y &\sim \mathcal{IW}(\boldsymbol{S}_0, \nu_0), \\ \boldsymbol{w}_i &\sim \mathcal{N}(0, \beta_i\boldsymbol{I}), \\ \beta_i &\sim \mathcal{IG}(\alpha_0, \beta_0) \end{aligned} \tag{46}$$

For $N \gg D$, a relative non-informative prior for $\boldsymbol{\Psi}$ can be specified by setting ν_0 to a small value, indicating a small number of virtual prior data. However, for scenarios with $N \approx D$, the posteriors of the $\boldsymbol{\Psi}$ becomes improper, and tricks like increasing the virtual sample count in the prior are required. This, in turn, makes the Wishart prior relatively strong, severely biasing the whole posterior.

In Equation 46, $\boldsymbol{w}_i$ denotes the column of $\boldsymbol{W}$ with a isotropic variance $\beta_i\boldsymbol{I}$. A hyperprior distribution for β_i is also determined. $\mathcal{IG}$ and $\mathcal{IW}$ are abbreviations for inverse-Gamma distribution and inverse-Wishart distribution, respectively. The priors for $\boldsymbol{\mu}$ and for the covariance matrices $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$ are conjugate priors. The prior for the projection matrix $\boldsymbol{W}$ is the so-called automatic relevance determination prior (ARD) [34]. It has been used for example in Bayesian Principal Component analysis [295] to determine the number of retained principal components automatically. Using values $\alpha_0 = 0.1, \beta_0 = 0.1, \nu_0 = p + 1, \boldsymbol{S}_0 = \boldsymbol{I}$ and $\sigma^2 = 1$ for the prior

and hyperprior parameters gives reasonably vague priors for a workable Bayesian data analysis.

The purpose of the ARD prior is to find low dimensional subspace for the canonical components. The number of canonical correlations is usually set to the minimum of the dimensionality of the two data sources. All of the canonical correlations are not significantly different from zero, so using ARD the true underlying dimensionality is found. The variance parameter β_i controls the magnitude of $\boldsymbol{w}_i$. If the dimensionality of the canonical subspace is less than the full dimensionality, the prior variance parameters for some columns of $\boldsymbol{W}$ go towards zero as do the actual elements of the vectors. The prior for the matrix $\boldsymbol{W}$ as a whole is not conjugate given ARD so the values need to be obtained component by component.

4 A new variant of Bayesian canonical correlation analysis to fuse gene expression and metabolomics data

4.1 Implementing group sparsity constraints into Bayesian canonical correlation analysis

The Bayesian probabilistic canonical correlation (BCCA) analysis presented in section 3.5.8 still has difficulties in modelling data sets having large p and small n. The standard BCCA model requires full rank covariance matrices, which prevents it from working with data sets having small number of samples compared to the number of variables. Existing Bayesian canonical correlation analysis have shown an improved performance on data sets with small sample size but they are still limited to small scale problems; in the experiments of several studies, the data dimensionality has been at most 10 [294, 296–298]. The difficulty arises from the full-rank covariance matrices having an inverse-Wishart prior because the estimation of these covariance matrices requires D(D+1)/2 parameter values to learn, D being the data dimensionality. For increasing D and fixed sample size N, the amount of data per element reduces rapidly, and inference slows down because the cost is cubic as a function of D. Inferring such large covariance matrices can be done by introducing a very strong prior assumptions such as restricting the covariance matrices to be diagonal. However, this would in turn produce serious bias in the results. In this section, a new Bayesian dependency model is introduced in which the covariance matrices for the data sets are estimated by matrices having a low rank [33]. The low rank matrix is constructed as an inner product of a matrix having a low number of columns. In contrast, some studies have proposed CCA variants that assume full sparsity for the projections/transformation matrices, thus decreasing the efficient number of parameters learned in the model [294, 297]. The former introduces sparsity to the transformation matrices via sparsity-inducing priors, and the latter by constructing an Indian Buffer process (IBP) prior for choosing the active elements [299]. The former approach still lacks the solution for identifying the shared and non-shared components, and the authors provide no empirical experiments with the full CCA model. The IBP-based model, in turn, retains the full-rank noise covariance, preventing efficient and robust inference for high-dimensional data [300, 301]. Another problem in the standard BCCA formulation is the lack of identifiability; the solution is found only up to an unknown rotation.

This chapter presents a new variant of Bayesian canonical correlation analysis which implements group sparsity constraints to the sample covariance matrices of the data sets [33]. The new model is similar to the model presented in Figure 10 which includes the data set-specific latent variables. Now the latent variables are not integrated out, which leads to a generative model having a plate diagram shown in Figure 12. If the data sets are assumed to be standardized, the probability

distributions assigned to the nodes are

$$\begin{aligned} \boldsymbol{z}, \boldsymbol{z}_x, \boldsymbol{z}_x &\sim \mathcal{N}(0, \boldsymbol{I}) \\ \boldsymbol{x}_j &\sim \mathcal{N}(\boldsymbol{W}_x \boldsymbol{z}_j, \boldsymbol{B}_x \boldsymbol{B}_x^T + \sigma_x^2 \boldsymbol{I}) \\ \boldsymbol{y}_j &\sim \mathcal{N}(\boldsymbol{W}_y \boldsymbol{z}_j, \boldsymbol{B}_y \boldsymbol{B}_y^T + \sigma_y^2 \boldsymbol{I}) \ . \end{aligned} \tag{47}$$

Actually, in Equation 47, it seems that the data set-specific latent variables have been integrated out. However, the covariance matrix of the feature-wise concatenation of the data, $\boldsymbol{\Psi}$, is a block diagonal matrix having $\boldsymbol{B}_x \boldsymbol{B}_x^T + \sigma_x^2 \boldsymbol{I}$ and $\boldsymbol{B}_y \boldsymbol{B}_y^T + \sigma_y^2 \boldsymbol{I}$ on its diagonal, and the data set-specific latent variables are needed to be included into the modelling to get estimates of the matrices $\boldsymbol{B}_x$ and $\boldsymbol{B}_y$.

If the data and the projection matrices $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ are concatenated as in Equation 45 and the covariance matrix $\boldsymbol{\Psi}$ is a block diagonal matrix having $\boldsymbol{B}_x \boldsymbol{B}_x^T + \sigma_x^2 \boldsymbol{I}$ and $\boldsymbol{B}_y \boldsymbol{B}_y^T + \sigma_y^2 \boldsymbol{I}$ on it's diagonal, the model in Equation 47 can again be presented in a simpler form:

$$\begin{aligned} \boldsymbol{z}, \boldsymbol{z}_x, \boldsymbol{z}_y &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ \boldsymbol{v} | \boldsymbol{z}, \boldsymbol{z}_x, \boldsymbol{z}_y &\sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z}, \boldsymbol{\Psi}), \ \boldsymbol{W} \in \mathbb{R}^{(p+q) \times d_z} \ . \end{aligned} \tag{48}$$

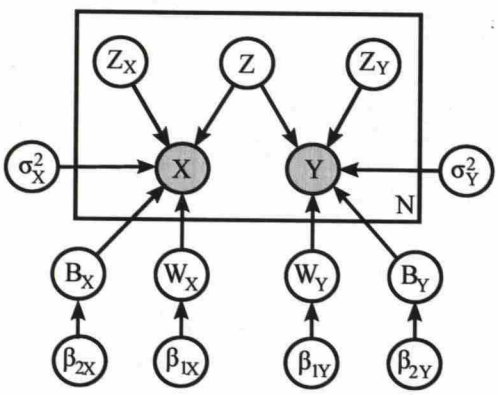


Figure 12: Probabilistic graphical model for canonical correlation analysis when introducing group sparsity constrains.

Priors for the model parameters are

$$\begin{aligned} \boldsymbol{w}_i &\sim \mathcal{N}(0, \beta_{1i} \boldsymbol{I}) \\ \beta_{1i} &\sim \mathcal{IG}(\alpha_{10}, \beta_{10}) \\ \boldsymbol{b}_{xk} &\sim \mathcal{N}(0, \beta_{2xk} \boldsymbol{I}) \\ \beta_{2xk} &\sim \mathcal{IG}(\alpha_{20}, \beta_{20}) \\ \boldsymbol{b}_{yl} &\sim \mathcal{N}(0, \beta_{2yl} \boldsymbol{I}) \\ \beta_{2yl} &\sim \mathcal{IG}(\alpha_{20}, \beta_{20}) \\ \sigma_x^2, \sigma_y^2 &\sim \mathcal{IG}(\alpha_0, \beta_0) \ . \end{aligned} \tag{49}$$

In Equation 48, $\boldsymbol{b}_{xk}$ is the kth column of $\boldsymbol{B}_x$, and $\boldsymbol{b}_{yl}$ is the lth column of $\boldsymbol{B}_y$. The priors for $\boldsymbol{w}_i$, $\boldsymbol{b}_{xk}$ and $\boldsymbol{b}_{yl}$ are all ARD priors. Virtanen et al. [33] present this model

in a following way: Feature-wise concatenation $\boldsymbol{v} = [\boldsymbol{x}, \boldsymbol{y}]$ leads to factorised CCA model presented in Equation 50. Instead of explicitly specifying three different sets of latent variables, namely one shared and two data set-specific components, a single set of components is used.

$$\begin{aligned} \boldsymbol{Z} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \\ \boldsymbol{v} &\sim \mathcal{N}(\boldsymbol{W}\boldsymbol{Z}, \boldsymbol{\Psi}) \end{aligned} \tag{50}$$

In Equation 50, $\boldsymbol{Z}$ includes the shared latent variable, as well as the data set-specific latent variables. The dimensionality of $\boldsymbol{Z}$ is $K_c = K + K_x + K_y$ and $\boldsymbol{\Psi}$ is a diagonal matrix that contains only the variances σ_x^2 and σ_y^2 on the diagonal in p and q copies. The structure of the transformation matrix $\boldsymbol{W}$ is as follows:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W_x} & \boldsymbol{B_x} & \boldsymbol{0} \\ \boldsymbol{W_y} & \boldsymbol{0} & \boldsymbol{B_y} \end{bmatrix}$$

This method is called the group-sparsity Bayesian canonical correlation analysis (gsCCA). The Bayesian canonical correlation analysis has reduced to a simplified factor analysis model with specific form of sparse structure for the linear projections $\boldsymbol{W}$. By inspecting the columns of $\boldsymbol{W}$, each column is either dense or has a very specific sparsity structure: either the elements corresponding to the first p rows of the matrix $\boldsymbol{W}$, the elements corresponding to the last q rows of the matrix $\boldsymbol{W}$, or both are zero. The last case implies a situation where the component is neither shared nor data set-specific, i.e. is pruned out.

The model structure presented above enables the learning of the covariance matrices in large p, small n setting, by automatically learning the low-rank structure of the covariance matrices $\boldsymbol{\Psi}_x$ and $\boldsymbol{\Psi}_y$. The automatic learning of the covariance structure is done by introducing automatic relevance determination priors [302,303] for the columns of the projection matrix $\boldsymbol{W}$. By integrating out the data set-specific latent variables, the model can be shown to be equivalent to imposing a low-rank assumption $\boldsymbol{\Psi}_x = \boldsymbol{B}_x\boldsymbol{B}_x^T + \sigma_2^2$ for the covariance matrices. This allows decreasing the computational demand and increases the amount of data per model parameter. However, the number of latent variables is increased three-fold. Nevertheless, this has the advantage that instead of merely capturing the correlation between the data sources, it produces a full factorization of the variation in data into shared and non-shared components. All this makes inferring the number of components in the model more difficult. For example, trying to simultaneously learn three component numbers with three separate ARD, and hence the structure of the matrix $\boldsymbol{W}$, is extremely sensitive to initialization. Correctly identifying the shared and non-shared components becomes tricky since the non-shared components can always be represented as the shared ones with an equal likelihood. The only information for identifying the components comes from the prior. For more discussion, see [33].

In contrast to the previous models, here the matrices $\boldsymbol{W}_x$ and $\boldsymbol{W}_y$ are modelled

row-wise, i.e. the ARD prior is determined for the rows of the projection matrices:

$$p(\boldsymbol{W}_x) = \prod_{d=1}^{p} \mathcal{N}(\boldsymbol{w}_{xd}|\boldsymbol{0}, \boldsymbol{\beta}\boldsymbol{I}) \tag{51}$$
$$p(\boldsymbol{\beta}) = \prod_{i=1}^{d_z} \mathcal{IG}(\beta_i|\alpha_0, \beta_0) \ .$$

Actually, Equation 51 still implies that each column of $\boldsymbol{W}$ has the corresponding precision β. Small values of α_0 and β_0 in a hyperprior for $\boldsymbol{\beta}$, shown in Equation 51, result in an ARD prior automatically driving unnecessary components to zero. If two grouped columns in the learned matrix corresponding to different data sets have a small β parameter, this component is completely pruned out of the model. If both groups have a high β, then this component is a shared component. If the β corresponding to one data set is large, and the β corresponding to other data is small, then the component is specific to either data set.

The prior distribution for $\boldsymbol{W}$ is

$$p(\boldsymbol{W}) = \prod_{d_1=1}^{p} \mathcal{N}(\boldsymbol{w}_{d_1}|\boldsymbol{0}, \boldsymbol{\beta}_1\boldsymbol{I}) \prod_{d_2=1}^{q} \mathcal{N}(\boldsymbol{w}_{d_2}|\boldsymbol{0}, \boldsymbol{\beta}_2\boldsymbol{I}) \ . \tag{52}$$

In Equation 52, $\boldsymbol{w}_{d1}$ is one of the p first rows of $\boldsymbol{W}$ and $\boldsymbol{w}_{d2}$ is one of the last q rows of $\boldsymbol{W}$.

gsCCA is inferred by a variational approximation method using formulas derived for a Bayesian PCA and Bayesian CCA where applicable [298, 302]. The factorised approximation of the posterior is presented in Equation 53. All model parameters are included in $\boldsymbol{\Theta}$. The different factors of the posterior are learned by updating them in cycles [33].

$$q(\boldsymbol{Z}, \boldsymbol{\Theta}) = q(\sigma_1^2)q(\sigma_2^2)q(\boldsymbol{Z}) \prod_{d=1}^{p+q} q(\boldsymbol{w}_d) \prod_{k=1}^{K_c} q(\boldsymbol{\beta}_1^{\boldsymbol{k}})q(\boldsymbol{\beta}_2^{\boldsymbol{k}}) \prod_{n=1}^{N} q(\boldsymbol{z}_n) \tag{53}$$

The underlying projection corresponding to the classical CCA solution can be identified for a Bayesian CCA with a post-processing step proposed by [291] and presented in Chapter 3.5.7. However, this extra step is essentially equivalent to solving the classical CCA problem, and hence many of the advantages of Bayesian treatment are lost. While the transformation can easily be applied to the maximum likelihood solution, it is unclear how the transformation works for the full posterior distribution. Virtanen et al have solved this unidentifiability problem by explicitly optimizing a variational lower bound with respect to an unknown rotation [33]. This procedure is similar to the approach presented by Luttinen et al which has dramatically improved convergence of variational approximation in factor analysis model [304]. Luttinen et al have achieved the speed up of the optimization by using proper parameterization of the posterior approximation, which allows joint optimization of its individual factors. For CCA, this approach does not only result in increase in computational speed of variational approximation; the optimization of

the rotation gives also the correct CCA projections, and helps separating the shared components from those specific to either data set [33, 304].

The trick that solves the rotational ambiguity comes after each round of variational updates. The variational approximation includes a separate parameter matrix $\boldsymbol{R}$ which is a linear transformation applied to $\boldsymbol{W}$. On the other hand, if $\boldsymbol{z}$ is multiplied by the inverse of $\boldsymbol{R}$, $\boldsymbol{W}^*\boldsymbol{z}^* = (\boldsymbol{W}\boldsymbol{R})(\boldsymbol{R}^{-1}\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$, the likelihood is invariant to $\boldsymbol{R}$. The rotation is inferred by maximizing the variational lower bound with respect to $\boldsymbol{R}$, or equivalently by minimizing the Kullback-Leibler divergence between the approximation $q(\boldsymbol{Z}, \Theta)$ and the prior $p_0(\boldsymbol{Z}, \Theta)$, shown in Equation 54. In Equation 54, the expectation is taken with respect to the transformed approximation $q_*(\boldsymbol{Z}, \Theta)$. For details of this optimization, see [304] and [33].

$$\arg \min_{\boldsymbol{R}} \langle \ln \frac{q_*(\boldsymbol{Z}, \boldsymbol{\Theta})}{p_0(\boldsymbol{Z}, \boldsymbol{\Theta})} \rangle_{q_*} \tag{54}$$

Given a fixed likelihood, the only way the variational bound can improve is by rotating the components so that the posterior $p(\boldsymbol{Z}, \boldsymbol{\Theta})$ better matches the factorial approximation $q(\boldsymbol{Z}, \boldsymbol{\Theta})$. Hence, maximizing the lower bound with respect to $\boldsymbol{R}$ equals forcing the model to find components that are a posteriori maximally independent of each other. This is analogous to the classical CCA solution requiring orthogonality of the components in the latent space. Hence, the maximasing $\boldsymbol{R}$ not only provides a deterministic choice for the rotation but also does it in a meaningful sense. For more discussion, see [33].

The gsCCA method has several good properties. It is computationally efficient and works for high dimensional data. In addition, the model separates the latent signals into shared and data set-specific. Virtanen et al have compared the performance of the original Bayesian CCA to gsCCA in finding the true shared components when varying the data dimensionality. The authors observed that the models have equal performance on low dimensional data but with high-dimensional data, the standard BCCA does not find the shared components correctly. The standard BCCA would require a full-rank covariance matrix, which is possible only in large n, small p settings. The new model, gsCCA, is able to extract the underlying model structure and is independent of the data dimensionality. Moreover, gsCCA was applied to neuroinformatics data set including 137 samples, 274 features in the other data set, and 28 features in the other data set. The gsCCA method applied to this data outperformed the standard multiple output linear regression model in the sense of mean-squared prediction error. One shortcoming of the model is that the rotation maximizes independence of the whole posterior approximations, which implies independence of both the latent variables $\boldsymbol{z}$ and the projection vectors. Ideally, a CCA model would only require independence over the latent variables. For more details, see [33].

In Chapter 3.2, it was shown how the regularization stabilizes the model parameters but introduces a small bias. Regularization in the Bayesian framework and using priors also biases the results. Priors are heavily used in gsCCA, so a question of the bias introduced by them rises. Inverse-Wishart priors for data covariance matrices in the standard Bayesian CCA are so called hard priors which biases the

results strongly if they are very informative. On contrast, the ARD prior is so called soft prior which adapts to the data complexity, and do not bias the results heavily.

4.2 Data for studying the role of the gene *RCD1* in *Arabidopsis thaliana* oxidative stress signalling

4.2.1 Description of the data

In the study of the role of the gene *RCD1* in *Arabidopsis thaliana* oxidative stress response, gene expression and metabolite data were measured in two genotypes: in the wild type Col-0 and in the mutant *rcd1*. Both genotypes were exposed to ozone starting at time point 0 h, and ending at time point 6 h. The data was measured at six time points starting from the ozone exposure: 0 h, 1 h, 2 h, 4 h, 8 h and 24 h. In addition to ozone treated plants, plants kept in clean air, i.e. in the control condition, were subjected to measurements as well. The experiment was performed three times, so there are 3 different measurement batches. The three different batches were measured during different seasons, so there might the variation due to that and other factors. In each batch and in each combination of genotype, treatment and time, 5-7 plants were analyzed. Separate MS measurements were obtained for each individual plant, but to obtain gene expression data, the biological material from all these 5-7 plants were pooled to get enough RNA material.

Gene expression was measured by custom made cDNA microarrays. In the gene expression analysis, common reference design [305] was used; to obtain a common reference, all biological material obtained in the experiment were pooled. Common reference and the sample of interest were labelled with the different dyes and hybridized on the same microarray slide. Dye-swap was performed thus giving two repetitions. For the MS experiment, the samples were prepared, and the amino acids derivatised using EZ-faast kit [306]. MS data were measures using Thermo Finnigen's LC-LTQ high-resolution MS instrument. Metabolites were identified using their retention time, mass and fragmentation pattern. The relative concentrations of amino acids were obtained by comparing their concentrations to three known inner standards. The data was collected using Xcalibur -program. The data was normalized with respect to the fresh weight of the samples. The relational metabolite levels for individual plants were averaged, as well as the repetitions obtained from the dye-swap. Together this gives 3(batches)$\times$2(genotypes)$\times$2(treatments)$\times$6(time points)=72 paired samples. A logarithmic transformation was applied to the metabolic data. Before logarithmic transformation, 1 was added to all metabolite data values to ascertain that after the logarithmic transformation, zero will be still zero. Logarithm transformation of the metabolite data takes away the effect that variance increases as the measurement level increases. One metabolite, ornithine, was removed from the data because it does not show any variance. As a result, the data include 27 metabolites.

The gene expression data were received as already pre-processed and normalized, so then relative intensity values were used as such. Recent reconstruction of *Arabidopsis thaliana* metabolic network built from the known reactions in the AraCyc

data base [307, 308], and the experimental data about the (ir)reversibility of the reactions was used to select a subset of genes among about 12500 genes analyzed in the microarray experiment [309]. Linear programming have been used to show that the model is capable of producing amino acids, nucleotides, starch, cellulose and lipids in the ratio needed for growth of heterotrophic *Arabidopsis thaliana* cell cultures [309]. All metabolite and reaction names in this metabolic model are AraCyc unique identifiers. The used model, however, did not include the names of the genes associated with the reactions, so those were queried from AraCyc database based on the reaction names using pathway tools using Common Lisp interface [310].

1164 genes associated with the reactions of Poolman's Arameta model were the basis of the gene subset selected for the analysis. This set includes genes related to the biological processes where the metabolites analyzed in this thesis participate [309]. In addition, 52 transcription factors which show differential expression at time points 1 h and 2 h in *rcd1* compared to the Col-0 either in control condition or under ozone exposure were included to the data set. The gene *RCD1* (AT1G32230) was also added. Together this makes 1217 genes. When choosing only those genes that are in the data, the number of genes decreases further. Moreover, only those genes having absolute fold change equal or larger than one, were included in the model. Finally there are 489 genes in the subset. Of the expression data of these genes, 1 % of the values were missing, so missing-value imputation was performed by an iterative version of principal component analysis. The method is called Non Linear Estimation by Iterative Partial Least Squares (NIPALS) [311]. This method has been the origin of partial least squares, and it allows performing PCA with missing data. NIPALS is implemented for example in the R package ade4 [312]. Finally the data was standardized to have zero mean and variance of one.

4.2.2 Pre-processing of gene expression and metabolite data using linear mixed effect models

Metabolite and gene expression data were both analyzed and pre-processed by linear mixed-effects models. The mixed-effects models were also used to address the significantly differential concentration levels of metabolites. The mixed-effects models resemble ANOVA type of methods in which the observed variance of a particular variable is partitioned into components attributable to different sources of variation. These different sources of variation are referred to as covariates, and in this work they are the genotype, treatment and time. Mixed-effects models contain both fixed effects and random effects. Fixed effects are covariates associated with an entire population, or with certain repeatable levels of experimental factors. On the other hand, random effects are associated with individual experimental units drawn at random from population. In other words, fixed effects assume that the data come from populations which may differ only in their means, whereas random effects describe a hierarchical structure of the data where the difference of different populations are constrained by the hierarchy. Random effects are additional error terms that account for correlation among observations within the same predefined group of samples. If there are more than one source of random variation, the model

is called a hierarchical model. For more details, see [313].

The linear mixed-effects model finds the differences in gene expression and metabolite levels due to genotype, treatment and time but it can also be used to remove the random effects such as the dye effect in gene expression studies, as well as the differences in three different measurement batches. For metabolite data, the unwanted random variation in different plants, or variation due to the MS technique could be modelled. There random effects are usually not of interest, and one would want to remove them from the data. For more details, see [313].

In linear mixed-effect models, both fixed and random effects are assumed linear. A linear mixed-effects model having a single level of grouping is presented as [313]

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \ldots, M, \\ \boldsymbol{b}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{\varepsilon}_i &\sim N(\mathbf{0}, \sigma\boldsymbol{I}) \ . \end{aligned} \tag{55}$$

In Equation 55, n_i dimensional observation vector $\boldsymbol{y}_i$ in group i is a linear function of p dimensional fixed effects vector $\boldsymbol{\beta}_i$, and $\boldsymbol{b}_i$ is a q dimensional random effects vector. Matrices $\boldsymbol{X}_i$ $(n_i \times p)$ and $\boldsymbol{Z}_i$ $(n_i \times q)$ are known fixed effects and random-effects regressor or contrast matrices, respectively. Gaussian error term $\boldsymbol{\varepsilon}_i$ has the dimension n_i and models the within-group error. The random effects $\boldsymbol{b}_i$ and the within-group errors $\boldsymbol{\varepsilon}_i$ are assumed to be independent for different groups, and to be independent of each other in the same group. In that case, the covariance matrix $\boldsymbol{\Sigma}$ of Gaussian random effects is diagonal.

Among the three measurement batches, replicates of the measurements for a given metabolite behave in a similar way but there is a shift in the overall concentrations. This is to be taken into account in a linear mixed effect model and, therefore, a random effect term b_v $(v = 1, 2, 3)$ for each measurement batch is added to the model. Also random effects for different plants are included, denoted as b_{vi} $(i = 1, \ldots, 5)$. The random effects are assumed to be independent of each other. The fixed effects include genotype, treatment and time and all combinations of these. The model is presented in Equation 56 for each metabolite. In this equation, C is the intercept term, α_k denotes treatment (k=1,2), β_g denotes genotype, $(g = 1, 2)$, and γ_t denotes time $(t = 1, \ldots, 6)$. The b_v is normally distributed with a diagonal covariance matrix, $b_v \sim N(0, \sigma_1^2)$, as well as b_{vi}, $b_{vi} \sim N(0, \sigma_2^2)$. The ε is the Gaussian error term; $\varepsilon_{gktdvi} \sim N(0, \sigma^2)$.

$$\begin{aligned} y_{gktvi} &= C + \alpha_k + \beta_g + \gamma_t + \alpha_k\gamma_t + \alpha_k\beta_g + \beta_g\gamma_t + \alpha_k\beta_g\gamma_t \\ &\quad + b_v + b_{vi} + \varepsilon_{gktvi} \end{aligned} \tag{56}$$

Linear mixed-effects model for expression of a gene is presented in Equation 57 where C, α_k, β_g and γ_t are same as in the Equation 56, and δ_d, $d = 1, 2$ denotes a fixed dye-swap effect. The b_v, $v = 1, 2, 3$, is the batch effect.

$$\begin{aligned} y_{gktdvi} &= C + \alpha_k + \beta_g + \gamma_t + \alpha_k\gamma_t + \alpha_k\beta_g + \beta_g\gamma_t + \alpha_k\beta_g\gamma_t + \delta_d \\ &\quad + b_v + \varepsilon_{gktdvi} \end{aligned} \tag{57}$$

The mixed effect model can be fitted using the maximum likelihood method. The fitted linear mixed effect model can and should be examined using graphical and numerical summaries. One graphical summary that should be examined routinely is the plot of the standardized residuals ε versus the fitted responses of the model. This plot is used to assess the assumption of constant variance of the residual error term. This plot is investigated to see whether there are systematic reduction or increase in the variance of ε as the level of the response decreases or increases. In addition, hypothesis test can be performed on the parameters learned by the linear mixed-effects model, and confidence intervals for the parameters can be calculated. The precision of fit and the significance of various terms in the model can be assessed, and the fitting of two different models to the data can be compared. A general method for comparing fit of the nested models is the likelihood ratio test [314]. A statistical model is called to be nested within another model if it represents a special case of another model. If L_2 is the likelihood of the more general model, and L_1 the likelihood of the restricted model, and the more general model fist better to the data, $L_2 > L_1$ and correspondingly $\log L_2 > \log L_1$ holds. The likelihood ratio test statistic (LRT) defined in Equation 58 will be positive in this case.

$$2\log(\frac{L_2}{L_1}) = 2[\log(L_2) - \log(L_1)] \tag{58}$$

The linear mixed-effects models were fitted to the data, and models analyzed using R-package nlme [315]. The raw metabolite data were pre-processed using linear mixed-effects models. Linear mixed-effects model was fitted to the data and p values corresponding to the differences of metabolites due to genotype, treatment and time were obtained. Problems in multiple comparison were avoided by adjusting the p values with Benjamini-Hochberg correction [67]. Finally, the metabolite data were normalized to have zero mean and unit variance.

5 Results

5.1 Correcting batch effect in the metabolite data by linear mixed-effects model

The linear mixed-effect model (Equation 56) was fitted to the metabolite data. The model fit to the data was assessed by investigating the distribution of residuals, i.e. the differences between the measured and predicted metabolite concentration levels. The residuals were normally distributed around zero in a similar way for all metabolites except serine. The model in Equation 56 was compared to a simpler model lacking the term b_{vi}. The likelihood ratio test indicated the more detailed model fitting better to the data, except for a metabolite pipecolic acid.

The random effects learned by the model (Equation 56) were subtracted from the original data. The effect of this pre-processing can be seen in Appendix A in Figure A1 where concentration levels for metabolite asparagine are plotted as a function of time for all combinations of genotype and treatment. Different colours red, blue and green correspond to different measurement batches. As can be seen, before pre-processing the measurements from different batches differ, whereas after pre-processing, they are more similar. However, the metabolite data are still very noisy.

5.2 Metabolites showing differential concentration levels due to genotype, treatment and time

The significantly differential concentration levels of metabolites due to genotype, treatment and time are obtained as a result from fitting the linear mixed effect model (Equation 56) to the metabolite data. The metabolites showing statistically significant (adjusted p value ≤ 0.05) differential concentration levels in the wild type Col-0 between control and ozone treatment are highlighted in Figure 13. Red colour indicates up-regulation, and green colour indicates down-regulation. The values in the table are logarithmic fold changse of the metabolite levels between compared conditions. As can be seen, the concentration levels of citrulline, cystathionine, glutamine and methionine are decreased at certain time points, whereas most of the metabolite levels increase from time point 2 h onwards. The levels of aromatic amino acids phenylalanine, tryptophan, and tyrosine increase due to ozone exposure. Valine, leucine, isoleucine, and lysine synthesised by the same metabolic pathway also show increased concentration in wild type plants due to ozone treatment.

The metabolites showing statistically significant differential concentration levels in the mutant *rcd1* between control condition and ozone exposure are shown in Figure 14. Citrulline and methionine show similar decreased concentration levels as the wild type. Moreover, the concentration levels of α-aminoadipid acid, glutamine, and methionine show similar differences in the mutant as in the wild type. Also the levels of arginine, histidine, isoleucine, leucine, phenylalanine, tryptophan, tyrosine and valine show similar higher expression but the fold changes are larger, and the higher levels occur earlier in the *rcd1* than in the wild type (isoleucine, leucine,

phenylalanine, glycine). Four additional metabolites show increased concentration in the *rcd1* compared to the wild type: α-aminobutyric acid, β-aminoisobutyric acid, γ-aminobutyric acid and pipecolic acid. Glycine does not show decrease at time point 2 h in the mutant plant as was observed for the wild type.

	time					
metabolite	**0 h**	**1 h**	**2 h**	**4 h**	**8 h**	**24 h**
alpha-aminoadipic acid						0.917
arginine						1.041
citrulline					-0.431	
cystathionine				-0.302		
glutamine						0.436
glycine			-0.753			0.906
histidine						0.824
isoleucine						0.708
leucine					0.528	0.863
methionine			-0.344	-0.568	-0.481	
phenylalanine			0.565	0.708	0.765	0.803
threonine					0.381	0.615
tryptophan				0.831	1.046	0.972
tyrosine			0.484	0.688	1.332	1.160
valine			0.281	0.348	0.382	0.416

Figure 13: The statistically significant differences in metabolite concentration levels due to ozone exposure in the wild type Col-0 (adjusted p value ≤ 0.05). The significantly up-regulated metabolites are highlighted red and the significantly down-regulated metabolites are highlighted green.

The changes of metabolite concentration levels between genotypes in control condition are shown in Figure 15. Several metabolite levels are differentially regulated in the mutant plant already in the control condition, and the differences change during the day. The metabolites α-aminoadipic acid, citrulline, glycine, phenylalanine and pipecolic acid show down-regulation, whereas alanine, arginine, glutamate, glutamine, lysine, proline and valine show up-regulation. Especially the proline levels are higher in the mutant plant in control condition than in the wild type.

The differential levels of metabolite concentrations between genotypes under ozone exposure are shown in Figure 16. The column corresponding to the time point 0 h is same as the corresponding column in Figure 15 because that is the sample obtained just before initiating the ozone exposure. Under ozone exposure the metabolites show differential expression from time point 2 h onwards. Citrulline is down-regulated at time point 24 h, later than in other comparisons. All other differentially regulated metabolites are up-regulated: α-aminobutyric acid, arginine, γ-aminobutyric acid, glycine, histidine, isoleucine, leucine, lysine, phenylalanine, proline, threonine, tryptophan, tyrosine and valine have all more elevated levels in the mutant *rcd1* than in the wild type under ozone exposure.

It is difficult to draw conclusion from the tables presented in Figures 13, 14, 15, and 16. They just show the significant differences in the metabolite concentration

	time					
metabolite	0 h	1 h	2 h	4 h	8 h	24 h
alpha-aminoadipic acid						1.258
alpha-aminobutyric acid						0.466
arginine						1.844
beta-aminoisobutyric acid						0.473
citrulline					-0.427	
gamma-aminobutyric acid				1.2003	1.0323	0.981
glutamine						0.450
glycine					0.5547	1.584
histidine						1.190
isoleucine			0.4105	0.6143	1.0290	1.896
leucine			0.4631	0.6300	1.4936	2.134
lysine					0.4239	0.922
methionine			-0.366	-0.584	-0.491	
phenylalanine		0.320	0.6384	1.3979	1.5242	1.959
pipecolic acid					1.5127	1.066
threonine					0.3187	0.835
tryptophan				1.0428	1.2623	2.122
tyrosine			0.4787	1.3237	1.7821	2.507
valine			0.2883	0.6565	0.7831	1.176

Figure 14: The statistically significant differences in metabolite concentration levels due to ozone exposure in the mutant *rcd1* (adjusted p value ≤ 0.05). The significantly up-regulated metabolites are highlighted red and the significantly down-regulated metabolites are highlighted green.

levels. The lists become more interesting when introducing to the analysis the expression values of genes associated with enzymes synthesizing and degrading these metabolites, as well as the transcription factors potentially regulating the expression of metabolism associated genes.

5.3 Significance of canonical correlations

5.3.1 Canonical correlations obtained by rrCCA

The significance of canonical correlations obtained by the ridge regression regularized CCA (rrCCA) were investigated using the test presented in Section 3.5.2. The training and test data were obtained by dividing the whole data into half. This division was performed 1000 times. For each division, 100 random samples $\boldsymbol{X}_{rand}$ and $\boldsymbol{Y}_{rand}$ were sampled from multivariate normal distribution with diagonal covariance matrix. Finally, this test resulted in 24 significant canonical correlations out of 27. This result is very likely overoptimistic.

The values of canonical correlations obtained by the rrCCA method are shown left in Figure 17. It can be seen that there is a clear gap between canonical correlations for canonical variate pair 5 and 6; between canonical variate pair 10 and 11; and between canonical variate pair 12 and 13. As suggested by Gonzaléz et

	time					
metabolite	0 h	1 h	2 h	4 h	8 h	24 h
alanine	0.3119					
alpha-aminoadipic acid	-0.783					
arginine	0.3059					0.397
citrulline				-0.584		
glutamate	0.3884					
glutamine	0.2934					
glycine				-0.700	-0.651	
lysine	0.4008					
phenylalanine		-0.367				
pipecolic acid	-1.059					
proline	0.6307	0.6646	0.706	0.4356	0.5926	0.668
valine	0.3920		0.311			0.373

Figure 15: The statistically significant differences in metabolite concentration levels in the *rcd1* in control condition compared to the Col-0 (adjusted p value ≤ 0.05). The significantly up-regulated metabolites are highlighted red and the significantly down-regulated metabolites are highlighted green.

al. [26, 27], the intrinsic dimensionality of the data set could be chosen to be 5, 10 or 12. As seen in the leftmost Figure of 17, the rrCCA method overfits badly as the smallest canonical correlations are not approaching zero as is expected.

5.3.2 Canonical correlations obtained by gsCCA

For the gsCCA method, the total number of shared and data set-specific components need to be set before training the model. To choose the correct number of components, gsCCA was run by varying the number of components between 2 and 50. The total number of components giving highest value for the lower bound was chosen. The lower-bound as a function of total number of components is shown in Figure 18. The highest lower bound is obtained using 19 components. The division of the components to shared and data set-specific components, and the final order of the shared components vary according to the initialization of the gsCCA algorithm; the results analyzed in this thesis consist of 13 shared components, one component specific to metabolite data, and 5 components specific to gene expression. The values of the canonical correlations for the 13 canonical variate pair are shown right in Figure 17. Because the initialization affects the results of gsCCA, 10 random initializations were performed with fixed number of components (19), and the solution giving the highest lower bound was chosen.

The significance of the canonical correlations obtained by gsCCA were addressed by investigating the posterior distribution of the canonical correlations. By using ancestral sampling, 500 samples from the model presented in Figure 12 were sampled.

	time					
metabolite	**0 h**	**1 h**	**2 h**	**4h**	**8 h**	**24 h**
alanine	0.312					
alpha-aminoadipic acid	-0.783					
alpha-aminobutyric acid						0.674
arginine	0.306			0.692		1.199
citrulline						-0.452
gamma-amino-butyric_acid				0.891	0.795	0.714
glutamate	0.388					
glutamine	0.293					
glycine						0.625
histidine						0.798
isoleucine				0.465	0.741	1.367
leucine			0.334	0.693	1.022	1.555
lysine	0.401					0.838
phenylalanine				0.494	0.473	1.114
pipecolic_acid	-1.059					
proline	0.631		0.500	0.825	0.559	0.631
threonine						0.347
tryptophan						1.268
tyrosine				0.620	0.495	1.445
valine	0.392		0.318	0.479	0.510	1.133

Figure 16: The statistically significant differences in metabolite concentration levels in the mutant *rcd1* under ozone exposure compared to the Col-0 (adjusted p value ≤ 0.05). The significantly up-regulated metabolites are highlighted red and the significantly down-regulated metabolites are highlighted green.

The histograms of the canonical correlations are shown in Appendix B in Figures B1 and B2. The canonical correlations are higher than zero with probability 0.95 for all canonical variate pairs except for the two last ones, the 12th and 13th canonical components. Therefore, the 11 highest canonical correlations were chosen as targets of further analysis.

5.4 Comparing performance between rrCCA and gsCCA

The performance of rrCCA and gsCCA were compared by investigating the canonical correlations obtained from the training and non-permuted validation data, as well as from permuted validation data using cross-validation. Moreover, the effect of optimization of the regularization parameters on the results obtained by rrCCA, or the optimization of the total number of components on gsCCA were addressed. The results of this comparison are shown in Figure 19. When using rrCCA, the non-permuted validation data canonical correlation coincide with the training data canonical correlation for the first canonical correlation (left in Figure 19). Some of the non-permuted validation set canonical correlations from fourth canonical correlation onwards are much lower than the training data canonical correlation, and the permuted and non-permuted validation data canonical correlations are around zero

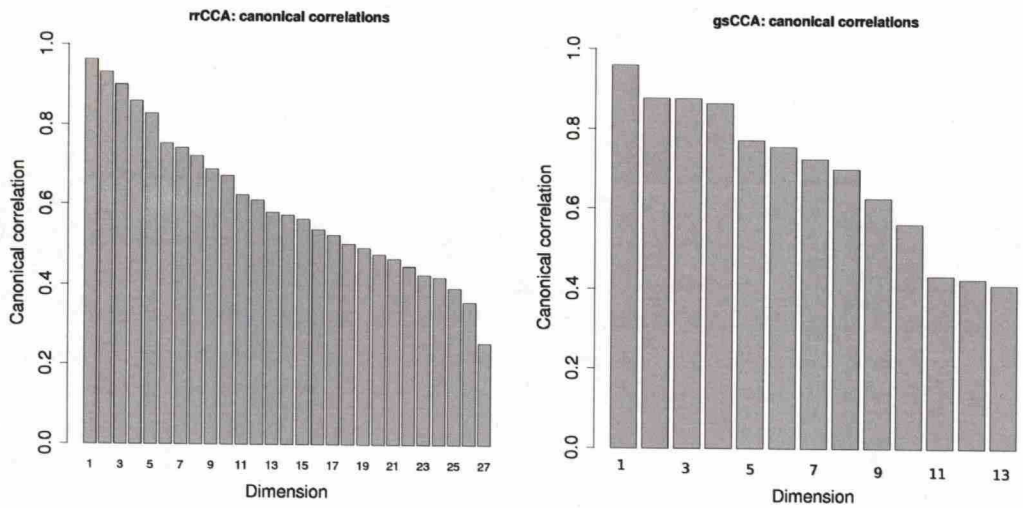


Figure 17: Canonical correlations obtained by the rrCCA and gsCCA methods.

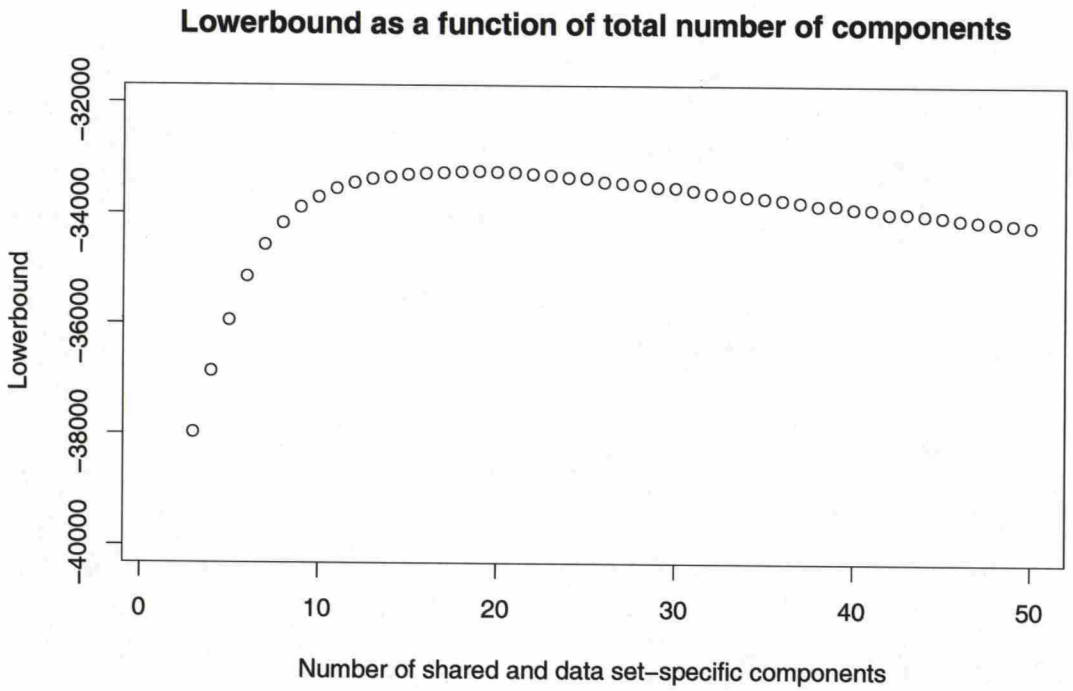


Figure 18: Lower bound as a function of the total number of components.

for the $11th$ canonical component. Therefore, the canonical correlations from the component 11 onwards are likely not significantly different from zero.

When the same experiment is performed using the gsCCA method, the correlations in the training data and non-permuted validation data are similar for the four highest canonical correlations, the fourth canonical correlation being equal in both training and validation data sets (right in Figure 19). The $7th$, $8th$ and $9th$

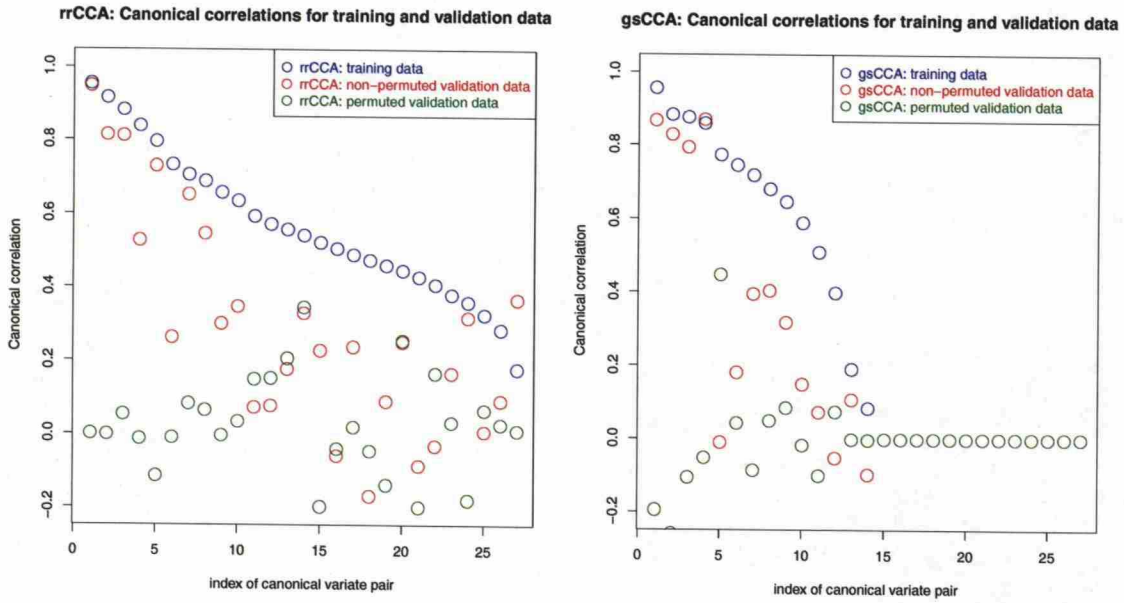


Figure 19: Training and validation set canonical correlations obtained by rrCCA and gsCCA when the model complexities are optimized.

canonical correlations between training and non-permuted validation data are also close to each other. However, the permuted validation data canonical correlation is higher than non-permuted canonical correlation already for the $5th$ canonical variate pair. Also the $12th$ and $13th$ permuted validation data canonical correlations are higher than non-permuted canonical correlations. From these figures and the results presented in the previous chapter, it can be concluded that the significance of the canonical correlations is better addressed by gsCCA than by rrCCA. Otherwise, there seems not to be much difference between the methods in obtaining the most highest canonical correlations. The results shown in Figure 19 are of course affected by the random division of the data into training and validation sets. In contrast to plotting only the means of canonical correlations of k training and validation data, the distribution of the canonical correlations could be better assessed by plotting also the distributions of the canonical correlations.

5.5 The amount of shared variance in data sets explained by canonical components

5.5.1 Variance explained by rrCCA

The variation in the data sets accounted for by the canonical components is investigated using squared intraset and interset canonical loadings. The average squared canonical loadings for the rrCCA method are shown in Figure 20. The first five components explain most of the variation in the data, although components 14 and 20 explain also decent amount of variation in the metabolite data. The intraset loadings in general are larger than interset loadings. The first five canonical compo-

nents, except the third one, explain more variation in the gene expression data; the third component explains more variation in the metabolite data. This plot clearly states that only the first five canonical components should be investigated more carefully. The cumulative distribution for the average squared canonical loadings are presented in Figure 21. The five first canonical components explain 62 % of the variation in the metabolite data and 70 % of the variation in the gene expression data based on the intraset squared canonical loadings.

The redundancy of the canonical variates obtained by the rrCCA method, i.e. the amount of shared variance between the two sets of variables in each canonical component, and the cumulative sum of redundancies over canonical components are shown in Figures 22 and 23, respectively. These figures result in similar conclusions as Figures 20 and 21. The cumulative sum of the redundancies of the first five canonical components is 0.54 for the canonical variates corresponding to the metabolite data, and 0.60 for the canonical variates corresponding to the gene expression data.

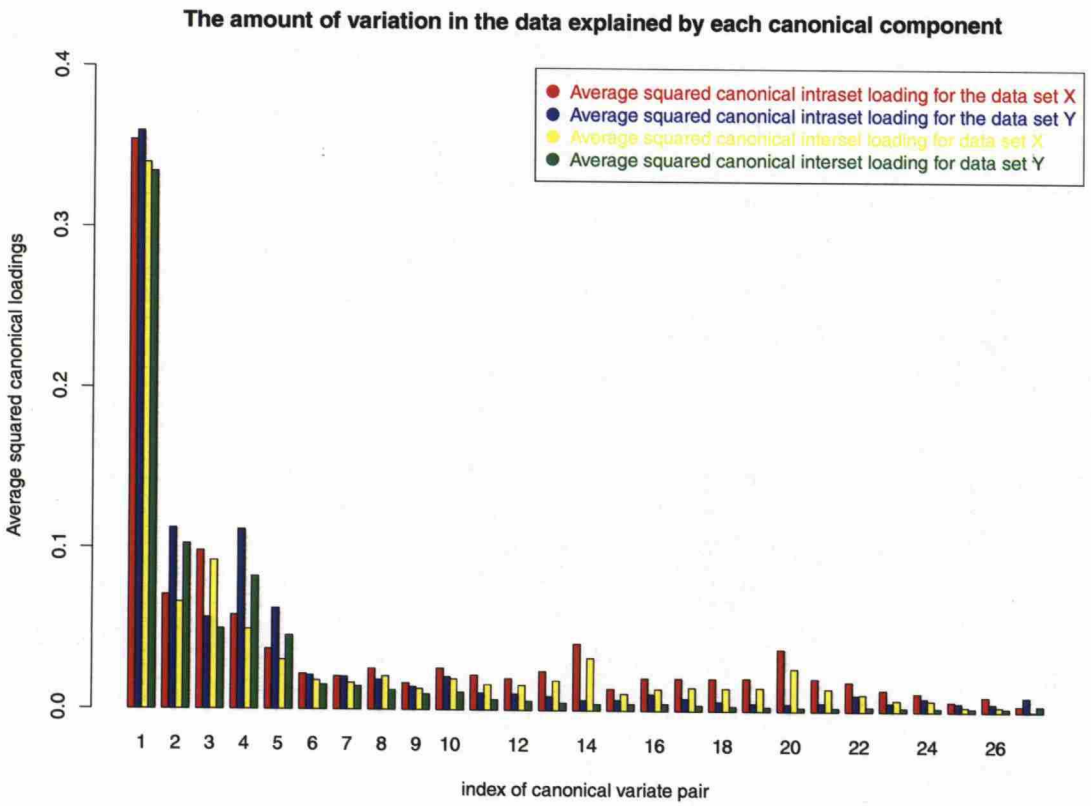


Figure 20: The proportion of variance in both data sets explained by the canonical variates obtained by rrCCA.

5.5.2 Variance explained by gsCCA

The average squared canonical intraset and interset loadings obtained for the gsCCA method in Figure 24 are somewhat different in comparison to values obtained using

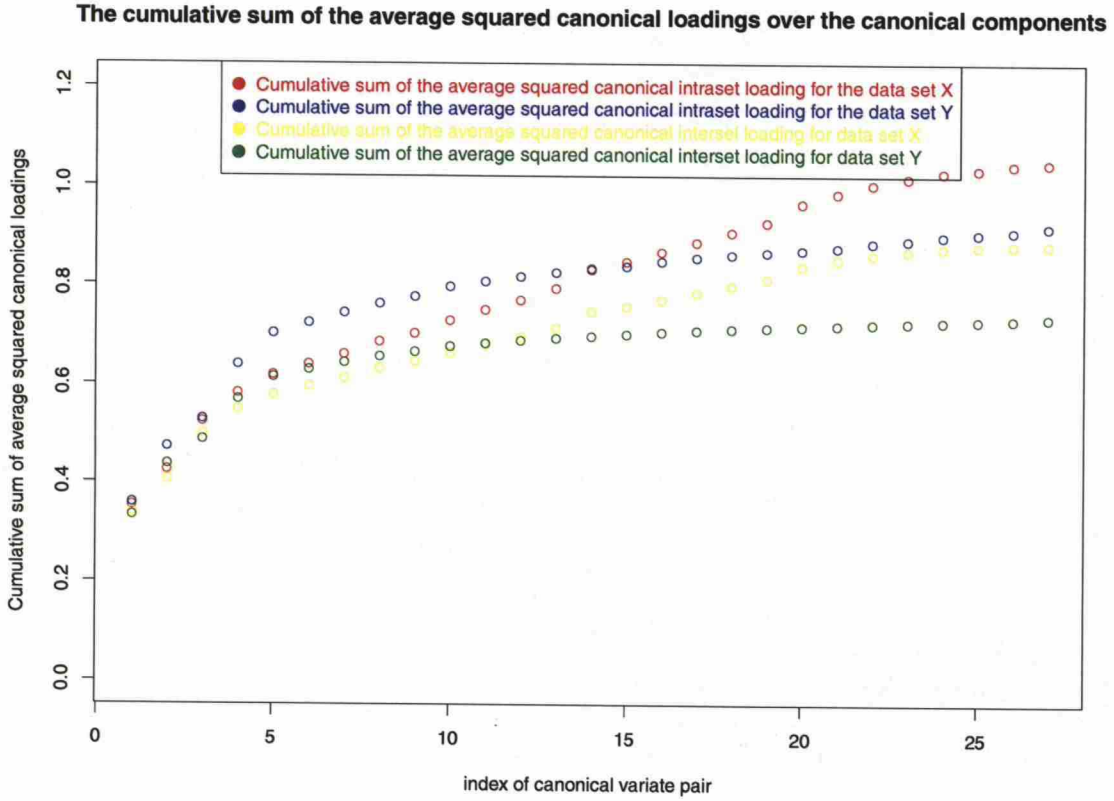


Figure 21: The cumulative sum of the proportion of variance in both data sets explained by the canonical variates obtained by rrCCA.

rrCCA. For gsCCA, the components 1, 2, 3, 4 and 8 include most of the variation, and component 5 explains still quite much variation in the metabolite data. Components 1 and 2 explain more variation in the metabolite data, whereas components 3, 4 and 8 explain more variation in the gene expression data. The Results indicate that the components 1, 2, 3, 4 and 8 should be studied more closely. The cumulative sum (Figure 25) of the average squared intraset canonical loadings for these components are 0.82 for both metabolite data and gene expression. The redundancies and the cumulative sum of redundancies are shown in Figures 26 and 27, respectively. The cumulative sum of the redundancies for components 1, 2, 3, 4 and 8 is 0.67 for the canonical variates corresponding to the metabolite data, and 0.62 for the canonical variates corresponding to the gene expression data.

When comparing the cumulative sums of the average squared canonical loadings and redundancies for the two methods, it can be concluded that the components obtained using gsCCA extract more variation from the two data sets, and the shared variation is also higher than obtained using rrCCA. High variation extracted from the metabolite data is a desirable property because in that case gene expression data with more variables does not dominate too much, and metabolism related shared processes are likely extracted better.

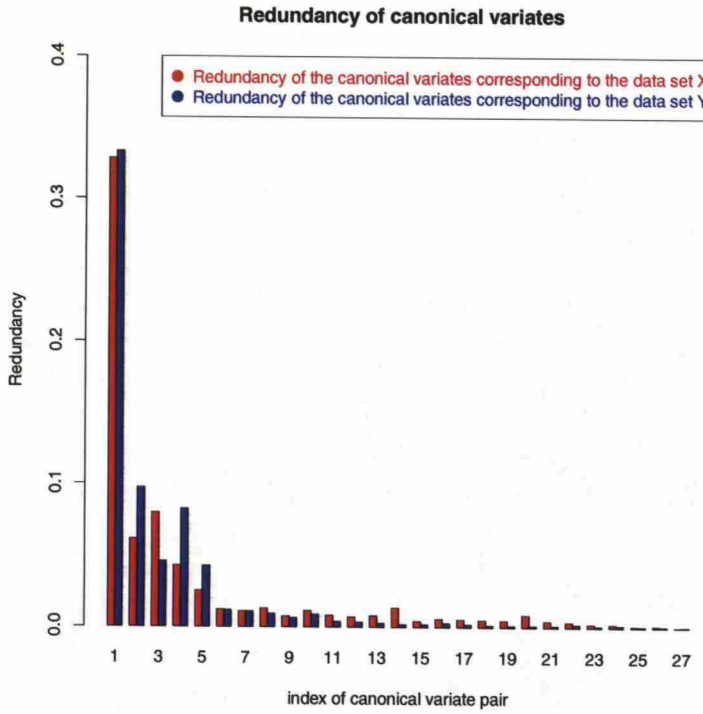


Figure 22: The redundancies of each canonical variate pair obtained by rrCCA.

5.6 Biological interpretation of canonical components

5.6.1 Interpretation of canonical components obtained by rrCCA

The samples analyzed by canonical correlation analysis can be projected onto the canonical components. For example, two dimensions of either canonical variate vectors can be chosen, and the samples can be projected on these two dimensions. In Figure 28, the samples projected on the first two dimensions of the canonical variate corresponding to the metabolite data obtained by the rrCCA method are plotted. Blue colour corresponds to the wild type and red colour corresponds to the *rcd1*. The labels control and O3 indicate the control and ozone treatment, respectively, and time point is marked at the end of the sample name. Figure 28 shows the ozone treated samples deviating from the control samples in the direction of negative x-axis corresponding to the first canonical component. Therefore, the first component likely contains the variation in the data caused by oxidative stress signalling: The ozone treated samples deviate from control until time point 8 h, and at time point 24 h the wild type returns close to the control state. However, the ozone treated mutant plants at time point 24 h still lie far away from the control samples.

In Figure 28, the second canonical component separates time points in control for both genotypes. This time dependent behaviour in control samples could be due to a circadian clock, a time-keeping mechanism that affects gene expression, and several regulatory mechanisms in plants such as environmental responsiveness

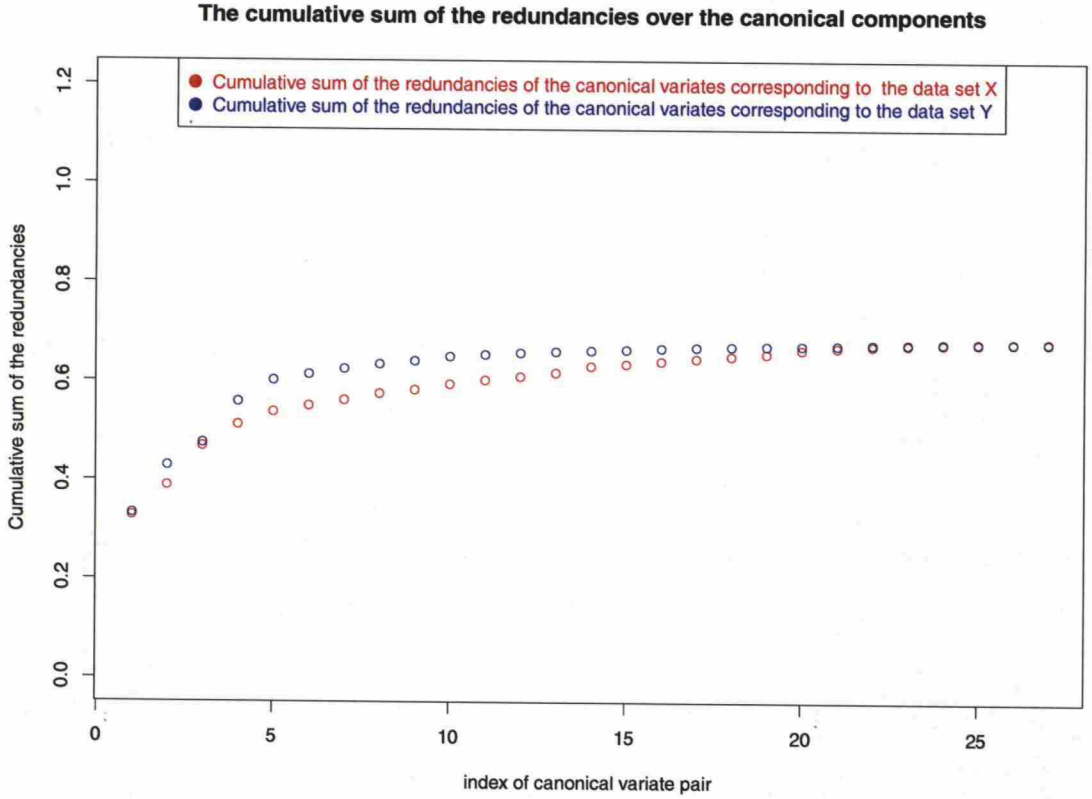


Figure 23: The cumulative sum of redundancies over the canonical variate pairs obtained by rrCCA.

to anticipate daily functions including light and temperature. To check whether the second component indeed models the circadian rhythm, the genes having a high canonical loading on any canonical components were compared to the list of circadian clock regulated genes reported in the work of Covington et al [316]. When investigating the percentage of differentially expressed genes having a high absolute canonical loading on components 1, 2, 3, 4 and 5, and belonging to the circadian rhythm regulated genes, the percentages were 29 %, 64 %, 7 %, 36 % and 0 %, respectively. The second component contains the highest percentage of the circadian rhythm regulated genes, so this component very likely models the circadian rhythm. As seen in Figure 28, in ozone treated plants, this rhythm is turned off while the plant starts to cope with the stress. However, the circadian clock affects also very likely the regulation of oxidative stress signalling, causing mixing of these two processes.

The samples projected on the first and third canonical components obtained using rrCCA are shown in Appendix C in Figure C1. The third component separates genotypes both in the control condition, as well as under ozone exposure. The figures of canonical variate pairs from 1 to 5 were investigated to assess the clustering and separation of the samples by the canonical components. The results are listed below for rrCCA method.

1. The first component separates ozone treated samples from control samples at

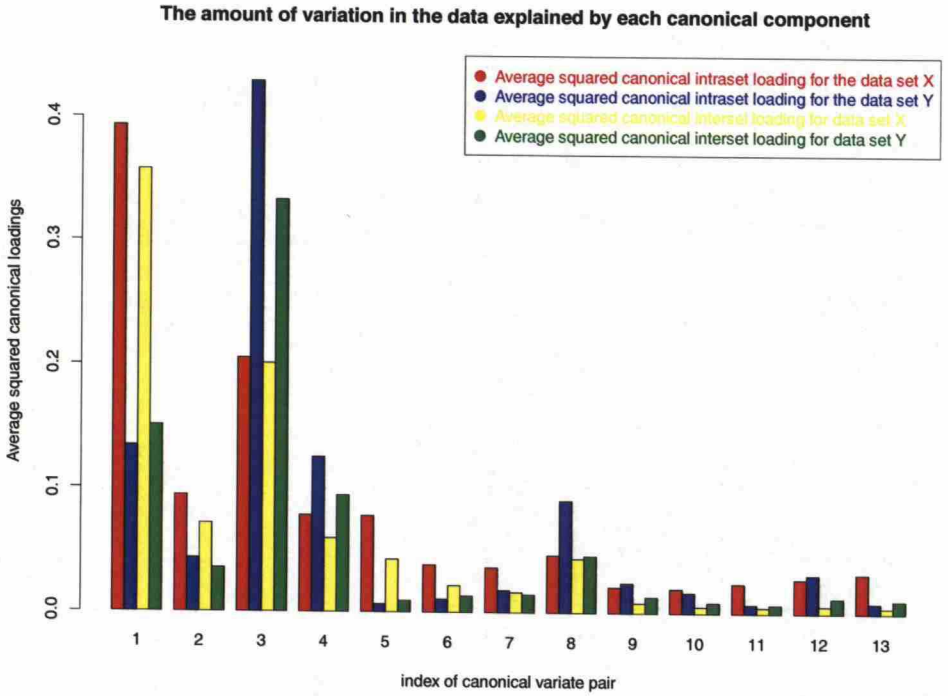


Figure 24: The proportion of variance in both data sets explained by the canonical variates obtained by gsCCA.

time points 2 h, 4 h, 8 h and 24 h. There is also a difference between genotypes in ozone at time points 8 h and 24 h. Time point 24 h for the *rcd1* in ozone is more deviated from control samples than for the wild type.

2. The second component clusters different time points of control samples. The variation in control samples is likely due to the circadian rhythm regulated genes.

3. The third component separates genotypes, both in control and under ozone exposure. This component also separates time point 1 h in ozone from control samples for the *rcd1*.

4. The fourth component separates ozone treated samples at time points 2 h and 4 h from time points 8 h and 24 h (early versus late regulation). There is also a slight difference between genotypes in ozone at time point 4 h.

5. The fifth component separates time points 1 h, 2 h, and 24 h from time points 4 h and 8 h under ozone exposure. Time point 1 h is more deviated from control samples for the *rcd1* than for the Col-0.

5.6.2 Interpretation of canonical components obtained by gsCCA

The samples projected on the first three dimensions of canonical variates obtained by the gsCCA method are shown in Appendix D in Figures D2 and D2. The sep-

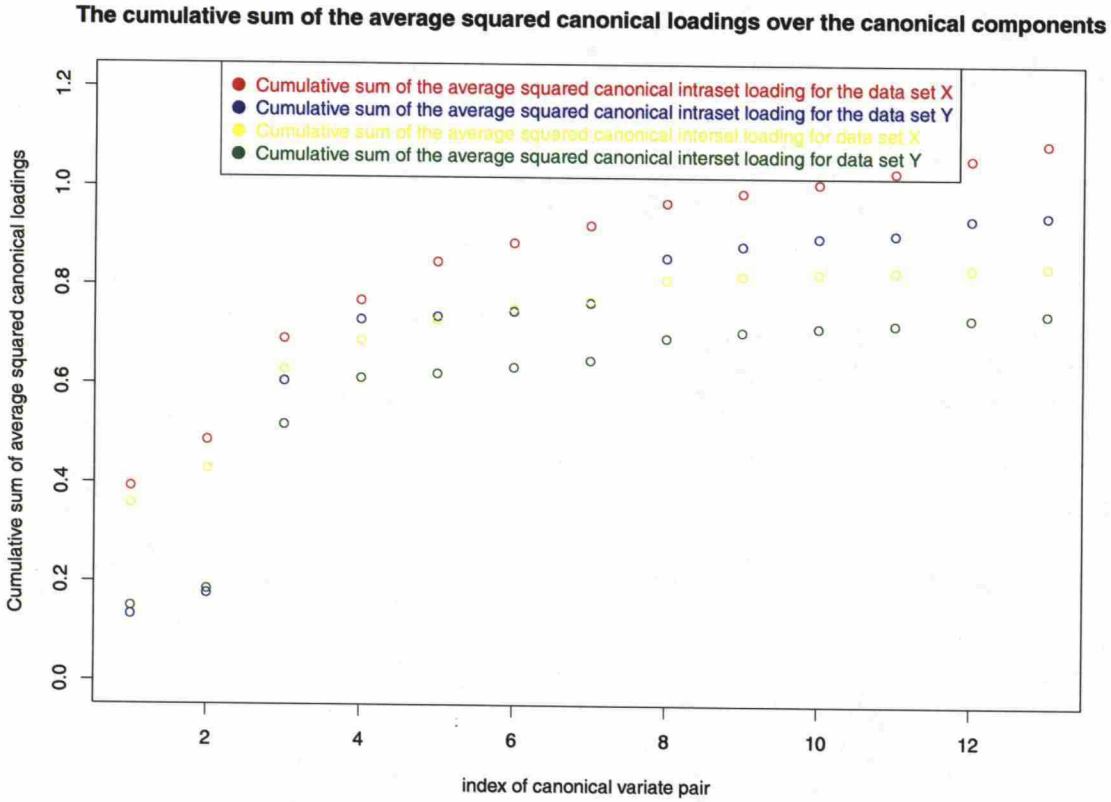


Figure 25: The cumulative sum of the proportion of variance in both data sets explained by the canonical variates obtained by gsCCA.

aration and clustering of samples by canonical components were studied, and the descriptions are listed below for the five interesting components. The percentages of circadian clock regulated genes on components were again investigated; the percentages were 25 %, 6 %, 29 %, 67 % and 17 % for the components 1, 2, 3, 4 and 8, respectively. Thus, the component 4 very likely models the circadian rhythm.

1. The first component separates ozone treated samples obtained at time points 8 h and 24 h from other samples. There is also a slight different between genotypes.

2. The second component separates genotypes. There is a slight difference between genotypes already at early time points in ozone. Samples for the *rcd1* in ozone at time point 24 h are clustered far away from the samples at time point zero, whereas for *Col*, the state of the plant at time points 0 h and 24 h is almost the same.

3. The third component separates ozone treated time points starting at time point 1 h from other samples.

4. The fourth component explains regulation of gene expression and metabolite

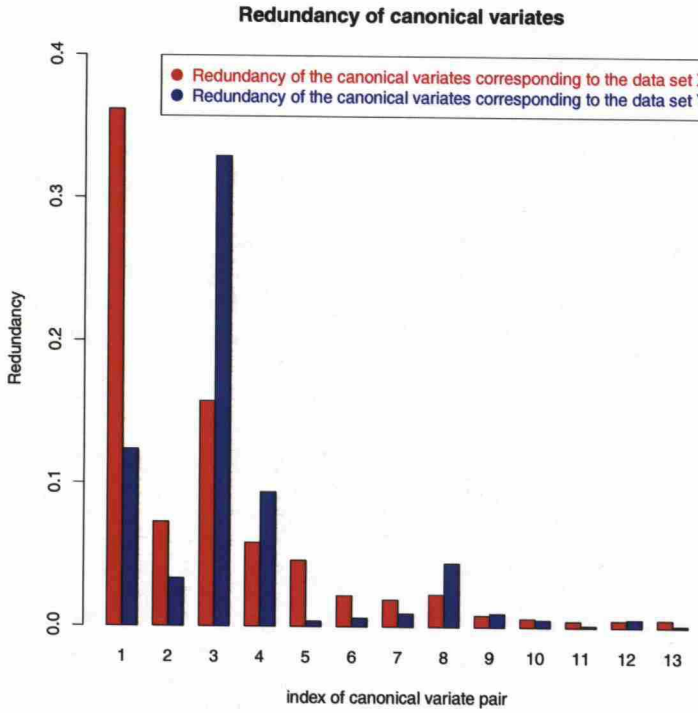


Figure 26: The redundancies of each canonical variate pair obtained by gsCCA

levels in control samples. This is likely due to the circadian rhythm regulated genes.

8. The *8th* component separates time points 1 h and 2 h from time points 0 h, 4 h, 8 h and 24 h under ozone treatment for both genotypes. However, time point 8 h for the *rcd1* in ozone does not fit to the pattern.

5.7 Genes and metabolites explaining the variation in canonical components

5.7.1 Results obtained by rrCCA

The genes and metabolites responsible for the clustering and separation of samples seen in Figures 28 and C1 can be addressed by investigating the canonical loadings. Canonical loadings can be visualized by canonical loadings plots; a canonical loadings plot showing the canonical loadings corresponding to the first two dimensions of canonical variates obtained by the rrCCA method is shown in Appendix E in Figure E1. In this plot, the coordinates for individual genes and metabolites are correlations between the original variables and the first two dimensions of canonical variates associated with the metabolite data set. In Figure 28, the first component separates the ozone treated samples from the control samples. Therefore, the genes and metabolites shown in Figure E1 having high or very low canonical loading on the first canonical component are responsible for the oxidative stress regulation. Only

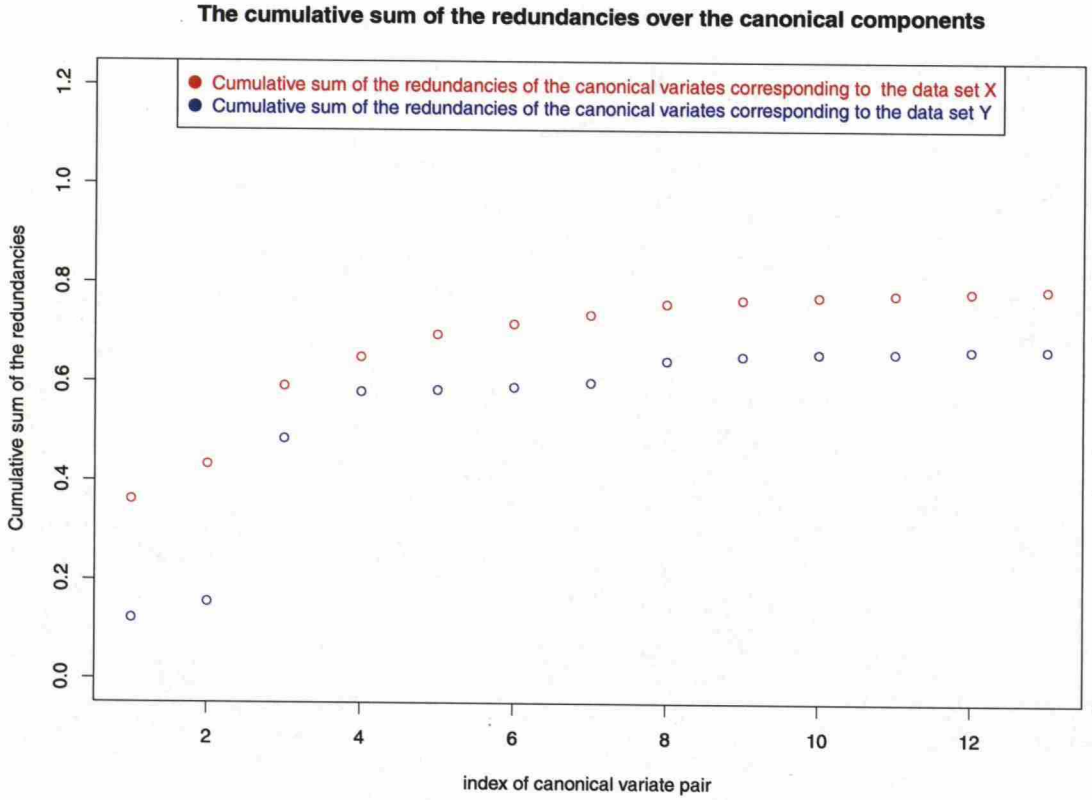


Figure 27: The cumulative sum of redundancies over the canonical variate pairs obtained by gsCCA

the genes and metabolites, whose distance from the origin of the figure is larger than 0.5, are plotted. This figure is difficult to read due to the large number of genes but the genes having high positive loading on the first component, in the right side of the plot, are down-regulated, and the genes having high negative loading on the first component, in the left side of the plot, are up-regulated during oxidative stress signalling. The side in which the up-regulated and down-regulated genes reside, needs to be checked each time.

Instead of plotting individual genes into correlation plots, the significantly enriched gene sets among high absolute canonical loadings can be plotted. This is shown for the first component of rrCCA in Appendix F in Figure F1 which plots metabolites, enriched biological processes gene sets, and transcription factors having a high canonical loading on the first and third component. In Figure F1, the metabolites and genes having high positive canonical loading on component 1 are up-regulated. In Figure F1, several amino acids, and a GO biological process class cellular amino acid and derivative metabolic process have a high canonical loading on the first component. This process is also associated with the third, fourth and fifth component (Figure J1 in Appendix J). The cellular amino acid and derivative metabolic process class includes genes that code enzymes which are involved, for example, in tryptophan, phenylalanine, proline, aspartic acid, methionine and flavonoid biosyn-

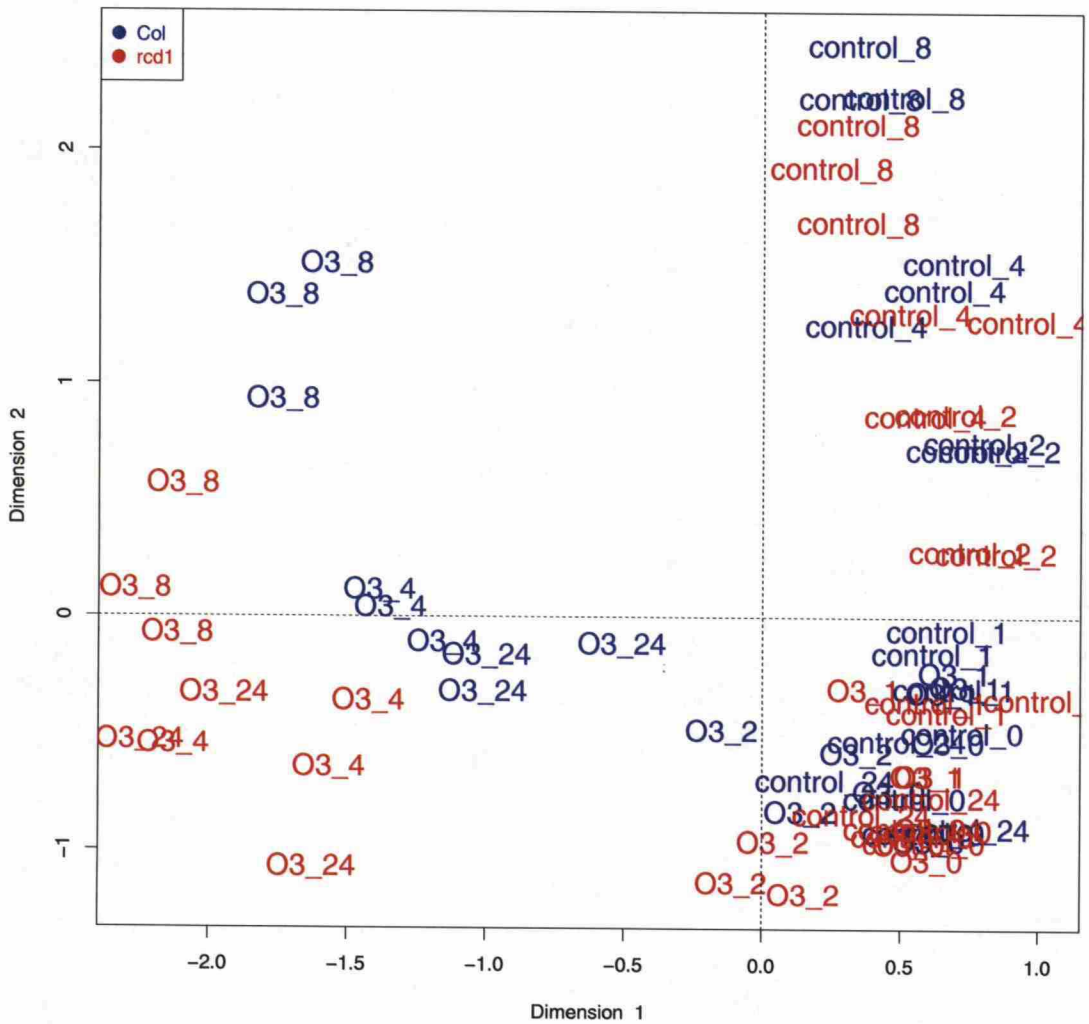


Figure 28: Samples projected on the first two dimensions of canonical variates corresponding to the metabolite data set. Canonical variates are obtained using rrCCA.

thesis and degradation pathways. The metabolites tryptophan, phenylalanine, proline, aspartic acid and methionine all have high canonical loading on component 1, as well as associations to other components (Figure J1). The genes belonging to the GO class cellular amino acid and derivative metabolic process, and involved in phenylalanine biosynthesis are plotted together with metabolite phenylalanine under ozone exposure in Appendix G in Figure G1. In this Figure, the up-regulated genes are up already at time point 2 h, and they return to their 0 h state latest at 24 h. The genes AT1G22410, AT1G48850, AT1G62960 and AT5G22630 are examples of early regulation explained by the fourth component. There are some differences in the gene expression between genotypes which could explain the difference in phenylalanine concentration observed between genotypes: 3-deoxy-7-phosphoheptulonate

synthase and arogenate dehydratase (AT5G22630) genes are up-regulated in the mutant *rcd1* at time points 4 h and 2 h, 4 h and 8 h, respectively, compared to the Col-0. Moreover, aspartate transaminase degrading phenylalanine is down-regulated at time point 4 h in the *rcd1* which could also explain more elevated phenylalanine levels in the mutant plant. 3-deoxy-7-phosphoheptulonate synthase, chrorismate synthase and aspartate transaminase are also involved in tyrosine, tryprophan and salicylic acid metabolism, so they might explain the differences in the levels of these metabolites between genotypes. Transcription factors shown in Figure F1, such as WRKY70, WRKY18 and NAP, having a high absolute canonical loading on component 1 could be up-regulating the genes belonging to the GO classes having high canonical loading on component 1. Despite the fact that several metabolic processes are up-regulated and/or down-regulated, the transcription factors are mainly up-regulated do to ozone exposure.

Chapter 2.5.1 presented metabolic pathways functioning in the oxidative stress signalling and producing various protective compounds against oxidative damage (Figure 7). The phenylalanine is the starting compound in many of these pathways. The first canoninical component obtained by rrCCA is associated to phenylalanine and genes metabolising it, thus validating that the first canonical component obtained by rrCCA models the oxidative stress signalling. Moreover, the biological process GO classes phenylpropanoid biosynthetic process, phenylpropanoid biosynthetic process and, as well as lignin biosynthetic process and lignin biosynthetic process are enriched in the forth and/or fifth canonical components, so these component very likely also model some aspects of the oxidative stress signalling.

5.7.2 Results obtained by gsCCA

Correlation circles obtained for gsCCA are shown for components 2, 3 and 4 in Appendix H in Figures H1 and H2. The most interesting regulation between genotypes in ozone occurs at time points 1 h and 2 h. Therefore, the $8th$ component obtained using gsCCA was analyzed. As seen in Appendix I in Figure I1, the components 2 and 8 separate the genotypes from each other, as well as the time points 1 h and 2 h in ozone from the other time points. For the mutant *rcd1*, the samples obtained at time point 1 h under ozone exposure are more deviated from the control samples than the corresponding samples of the wild type.

The genes and metabolites having high absolute canonical loadings on components 2 and 8 are shown in a correlation circle presented in Appendix I in Figure I2. The metabolites having high negative canonical loadings on the $8th$ component are serine and aspartic acid. However, these metabolites do not show statistical significant difference in any of the comparisons (Figures 13, 14, 15, and 16). Observing genes related to serine or aspartic acid metabolism having similar expression pattern as the concentration pattern of the metabolites could bring more statistical strength to the pattern observed for these metabolites. The biological process GO classes enriched on the $8th$ component are shown in Figure I3. It can be seen that a GO class cellular process is enriched in this component. The genes belonging to this class together with a metabolite aspartic acid are plotted in Figure I4 un-

der ozone exposure. These four genes are strongly up-regulated at time points 1 h and 2 h, but return after that to their initial state (except gene AT4G14880 which is down-regulated at time point 8 h). The aspartic acid concentration levels are slightly up at time points 1 h and 2 h in the Col-0, a pattern that is missing in the *rcd1*. In contrast, aspartic acid levels in the *rcd1* degrease at time point 8 h. The genes AT3G23250 and AT4G14680 also have differences in their expression profiles between genotypes, which might explain the changes in aspartic acid concentration levels.

5.8 Visualization of the gene sets associated with the canonical components by eye diagrams

5.8.1 Eye diagrams for the results obtained by rrCCA

Significantly differentially expressed gene sets associated with different canonical components are further visualized using eye-diagrams. The significantly differentially expressed gene sets were determined based on the canonical loadings of the corresponding genes. The enrichment was tested based on area under precision-recall curve. The p values indicating the statistically significant gene sets were adjusted using Benjamini-Hochberg method. Gene sets having adjusted p value ≥ 0.05 are considered significant, except for the biological functions GO classes significance level 0.01 is used to restrict the number of enriched gene sets.

The eye diagram shown in Appendix J in Figure J1, Z1, Z3, Z4 and Z5 denote canonical components obtained by the rrCCA method. The second component associated with the circadian rhythm-regulated genes is not included because it would introduce too many enriched classes. The metabolites having high absolute canonical loading on the components are connected to the components by colour-coded curves. The width of the curve reflects the strength of the association. In Figure J1 on the right, significantly enriched GO classes belonging to the GO category biological processes on the different components are visualized. Figures J2, J3 and J4 show the metabolic functions and cellular compartmentalization GO categories and AraCyc pathways, respectively. Gene set names coloured by red indicate that most of the genes belonging to that gene set have positive canonical loading, whereas blue gene sets include genes having mostly negative canonical loadings. If two canonical components are associated with a particular gene set, and the signs of the average canonical loadings are different in different components, the name is coloured green.

The first component obtained by rrCCA separates ozone treated samples from control samples. As was stated in the previous chapter, the down-regulated processes during ozone stress signalling have a high positive average canonical loading on the first component, so their names are coloured red in Figure J1. The gene sets having high negative canonical loading on average are up-regulated, and their names are coloured blue. Several gene sets having a positive average canonical loading on the component 1 are associated with chloroplast and plastid, and are therefore down-regulated (green gene sets in Figure J3. Chapter 2.5.1 mentioned that the photosynthetic processes occurring in chloroplast are down-regulated due

to ozone-signalling. Thus, the first component correctly finds oxidative stress signalling related processes. Figure J1 shows that the first component is also associated with several amino acid and organic acid related pathways. The gene sets associated with the first component contain very many genes, and usually some of them are up-regulated whereas others are down-regulated due to ozone exposure. Therefore, the colour of the gene set name is not always very informative.

The third component separates genotypes, both in control condition and under ozone exposure. Processes having average positive canonical loading on this component in Figure J1 are up-regulated in the mutant. This component includes several transcriptional regulatory and response processes, as well as metabolic and developmental processes. In addition to metabolism related cellular compartments enriched in this component, there is also nucleus (Figure J3); this implies that several transcription factors are responsible for the differences depicted by the third component. This can be also seen in Figure J2 where metabolic functions nucleic acid binding, transcription factor activity, DNA binding and transcription regulator activity are enriched and up-regulated. Figure J4 shows that AraCyc pathways glycolysis I and II and gluconeogenesis are associated with the third component. Of the metabolites, proline has a high canonical loading on this component.

The fourth component separates time points 2 h and 4 h and from time points 8 h and 24 h under ozone exposure. In Figure J1, the gene sets having negative canonical loading on average (blue colour) on the fourth component are up-regulated at time points 2 h and 4 h (sometimes also at time point 8 h) and gene sets having positive canonical loading on average (red colour) are down-regulated at time points 2 h and 4 h. This component contains metabolic processes, several response processes, developmental processes, transferase activity and membrane associated processes (Figures J1, J2 and J3). The regulation of these processes occurs at time points 2 h and 4 h. Of the metabolites, alanine, α-aminoadipic acid and β-aminoisobutyric acid are highly associated with this component.

The fifth component separates time points 1 h, 2 h and 24 h hours from time points 4 h and 8 h under ozone exposure. The gene sets having positive canonical loading on average (red colour) on the fifth component in Figure J1 are up-regulated at time points 1 h and 2 h, and the expression goes to zero at time points 4 h and slightly increases again at time point 24 h. The gene sets belonging to this group are various responses to stimulus and lignin, as well as phenylpropanoid biosynthetic process. The component is highly associated with the metabolite glutamate.

Several components are associated with the same gene sets, especially the metabolic processes associated with the first component. The gene sets associated with several components, one of them being the third component separating the genotypes, might be of interest. In Figure J1, for example, cellular aromatic compound metabolic process, responses to various stresses, and regulation of processes are associated with both the third and the fifth component. By investigating these shared gene sets, it might be found, how genotypes are differently regulated at time points 1 and 2 h.

5.8.2 Eye diagrams for the results obtained by gsCCA

Eye diagrams were also used to visualize the results obtained by gsCCA. The first component obtained by gsCCA separates ozone treated samples from others at late time points. The eye diagram of biological processes in shown in Appendix K in Figure K1. The gene sets associated with the first component with a negative canonical loading on average are up-regulated at time points 4 h, 8 h and 24 h, and gene sets with a positive canonical loading on average are down-regulated at the same time points. This component is associated with very many biological processes GO classes such as amino acid and organic acid metabolism, signalling, responses to various stresses and hormone biosynthesis and response processes, and it is associated with all enriched cellular compartments GO classes (Figure K3). Of AraCyc pathways, the first component is associated with glyoxylate cycle and TCA cycle (Figure K4). It can be concluded that the late ozone response is very dominant in the data.

The second component obtained by gsCCA separates genotypes. The gene sets having high average positive canonical loading are up-regulated in the mutant plant, sometimes already in the control condition. The second component is very similar to the third component obtained by rrCCA, with GO classes including regulatory and response processes. Also other eye diagrams (Figures K2, K3, K4) show the enriched gene sets for the second component are very similar to those associated with the third component obtained by rrCCA.

The third component separates ozone treated time points starting at time point 1 h. In Figure K1, the gene sets having positive average canonical loading on the third component are up-regulated due to ozone exposure, whereas gene sets having negative canonical loading are down-regulated. This component is associated with amino acid metabolism, especially histidine biosynthesis process, and heterocycle biosynthetic process, as well as some stress responses. AraCyc pathways in Figure K4 shows that glycolysis I and II are down-regulated, as well as sucrose degradation to ethanol and lactate. The third component is somewhat similar to the first component obtained by the rrCCA method but it includes stress responses and specific amino acid related process (histidine biosynthesis).

The $8th$ component separates time points 1 and 2 from other time points in ozone. In the eye diagrams, the gene sets having positive average canonical loading on the $8th$ component are up-regulated at time points 1 and 2 h. This is the early ozone response, and includes regulation and biosynthetic processes, as well as responses to chitin, carbohydrate and organic substance stimulus. Metabolic functions shown in Figure K2 include, for example, increased transcription factor activity. The component is, however, not associated with cellular compartment nucleus (Figure K3).

In eye the diagrams obtained for the results obtained by gsCCA, the associations of components to metabolites are stronger, which can be seen from the width of the curves connected to metabolites. Again the shared gene sets between genotype separating component 2, and components 3 or 8 are the most interesting results.

6 Conclusion and future work

In bioinformatics and systems biology, high-dimensional data sets obtained by high-throughput techniques pose several challenges to data analysis. Fusing the different types of omics data sets observed on the same experimental units has gained wide interest. Data fusion could be performed by seeking the variation that is common to several data sets. This can be done using canonical correlation analysis. In this thesis, the performance of a new variant of Bayesian canonical correlation, BCCA with group sparsity (gsCCA), is compared to the performance of the classical ridge regression regularized canonical correlation analysis (rrCCA). The gsCCA method is claimed to be the first applicable version of Bayesian CCA to data having a high dimensionality but a small sample size, so it is expected to perform better than rrCCA. The suitability of the methods to find regulatory relationships between transcripts and metabolites was addressed by applying them to data from a study of plant oxidative stress signalling where *Arabidopsis thaliana* was exposed to ozone, and paired time series gene expression and metabolite data were obtained both under ozone exposure and in control condition. The performance of the both methods is very similar; rrCCA and gsCCA both are able to find large and significant canonical correlations between the data sets, and the canonical components corresponding to these canonical correlations explain interesting biological variation. The gsCCA method was shown to explain more shared variation between the data sources than rrCCA, especially the variation in the metabolite data. The biological interpretation of the canonical components obtained differ, so gsCCA may find additional shared variation in the data sets that rrCCA is not able to find. Both methods are shown to find interesting relationships between transcripts and metabolites.

The gsCCA method has some desirable properties compared to rrCCA: First, the optimization of the regularization parameters for rrCCA takes several days, whereas the gsCCA algorithm is run on few minutes. Moreover, the optimization of the regularization parameters gives different results when the optimization is repeated. This is due to the random division of the data into training and validation set in the cross-validation procedure. The gsCCA method has the advantage that all data is used to learn the complexity of the model. Second, assessing the significance of canonical correlations is more straightforward when using gsCCA. Third, canonical components obtained using gsCCA explain more shared variation in the data sets than those obtained by rrCCA. Therefore, gsCCA is an attractive method for high-throughput data fusion. However, the analyses presented in this thesis have several limitations. The real biological data was very noisy, especially the metabolite data. Moreover, the difference between the dimensions in metabolite data and gene expression data was quite large. It was observed that gsCCA also overfits to the data. The effect of the noise level in the data, the unbalanced number of features, and the dimensionality of the data sets on the performance of the methods could be addressed using toy data. Better visualization methods of the results are needed as, for example, the canonical loadings plots are difficult to read if the data contain hundreds of variables. The selection of the subset of genes for the analysis based on metabolic network model introduces genes that are not related to the metabolites

analyzed in the work, and which just bring additional non-relevant variation to the data set and increase the number of false positives. The subset of genes could have been selected based on other metabolic network models, or by a detailed knowledge of the biological processes under study. Also the set of transcription factors included in data could have been chosen better. The first canonical component obtained by gsCCA separates the time points 8 h and 24 h between genotypes under ozone exposure. This component probably models the programmed cell death. However, the TFs showing differential expression at time points 1 h and 2 h were selected to the analysis. The TFs showing differential expression at time points 4 h and 8 h could explain the variation seen in the first canonical component obtained by gsCCA.

There are some problems associated with the canonical loadings. Canonical commonality analysis has been suggested to be used for the interpretation of the canonical components. It was not used in this thesis as the calculation of canonical commonalities is tedious, and existing software packages perform commonality analysis only for the non-regularized classical canonical correlation analysis. The gsCCA method could be developed further. One potential improvement could be the introduction of variable selection [317–319] using LASSO [37, 320, 321], or even better, Elastic-net regularization [36, 38, 286, 322–324]. These regularization methods would prune out the variables which do not contribute to the relevant variation among the samples. This would lead to sparse transformation matrices, and make the interpretation and visualization of the results much easier as there would be no need to investigate the parameter values corresponding to all features. Priors corresponding to LASSO or elastic-net has been introduced [325, 326], and might be introduced to the Bayesian gsCCA. The gsCCA method could be extended so that it can be applied to more than two data sets. For example, CCA applicable to four data sets could fuse data on different levels of biological regulation such as expression of transcription factors, expression of metabolism related genes, protein levels, and metabolite levels. The method should be applicable also to discrete data, for example SNP profiles. This could be achieved using optimal scaling of the variables [327–330], leading to non-linear canonical correlation analysis [331]. gsCCA should be applicable to data obtained by next-generation sequencing techniques as it is and will be one of the main high-throughput method used. Biological processes which change due to the genotype effect, as well as the ozone stress effect could be better found by the BCCA developed by Huopaniemi et al which forces the model to find particularly interaction effects of two covariates [32]. Including this property into gsCCA could also be one way to develop gsCCA further.

The modelling and visualization of the results could be performed in the context of biological background information. There are methods which can simultaneously visualize gene expression and metabolites in the context of biological pathways [170, 183, 332–335] but no method exists which can visualize canonical loadings in the context of biological pathways. Nevertheless, rrCCA and gsCCA were able to find potential interesting regulatory relationships between genes and metabolites. It is notable that there are no protein degradation processes among enriched gene sets because the changes in amino acid levels may result from degradation of proteins to amino acids. When adapting to the stress, the plant needs to synthesise

new proteins and the amino acids required for the synthesis may be obtained from other less necessary proteins by degrading them leading to increased concentration of amino acids. The situation is quite an opposite; several metabolic processes directly producing and consuming metabolites analysed in this work are enriched. Of course, the results obtained using current methods could be further investigated to find more interesting relationships between metabolites and genes. For example, one could easily project the genes not included in the analysis described in this thesis to the canonical components obtained by the two methods, and perform gene set enrichment analysis on them to find more enriched biological processes associated with the components. One of the future directions would also be to find common sequence motifs at the promoters of the gene sets enriched to find potential transcription factors regulating the genes, and inteacting with the mutated protein RCD1.

References

[1] J. E. Cohen, "Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better," *PLoS Biol*, vol. 2, no. 12, pp. e439+, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1371/journal.pbio.0020439.

[2] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, no. 6761 Suppl, pp. C47–C52, 1999. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/35011540.

[3] H. Kitano, "Systems Biology: A Brief Overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.1069492.

[4] H. Kitano, "Computational systems biology," *Nature*, vol. 420, no. 6912, pp. 206–210, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nature01254.

[5] U. Sauer, M. Heinemann, and N. Zamboni, "Getting Closer to the Whole Picture," *Science*, vol. 316, no. 5824, pp. 550–551, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.1142502.

[6] A. E. Oostlander, G. A. Meijer, and B. Ylstra, "Microarray-based comparative genomic hybridization and its applications in human genetics," *Clinical Genetics*, vol. 66, no. 6, pp. 488–495, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1399-0004.2004.00322.x.

[7] J. M. Johnson, S. Edwards, D. Shoemaker, and E. E. Schadt, "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments," *Trends Genet*, vol. 21, no. 2, pp. 93–102, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/j.tig.2004.12.009.

[8] A. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nrg1272.

[9] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, "A vision for the future of genomics research," *Nature*, vol. 422, no. 6934, pp. 835–847, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nature01626.

[10] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature*, vol. 461, no. 7261, pp. 218–223, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nature08454.

[11] M. J. Herrgård, V. Lee, B. andPortnoy, and B. Palsson, "Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae," *Genome Research*, vol. 16, no. 5, pp. 627–635, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1101/gr.4083206.

[12] T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin, "A genome-scale computational study of the interplay between transcriptional regulation and metabolism," *Mol Syst Biol*, vol. 3, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/msb4100141.

[13] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.278.5338.680.

[14] T. Shlomi, M. N. Cabili, M. J. Herrgård, B. O. Palsson, and E. Ruppin, "Network-based prediction of human tissue-specific metabolism," *Nat Biotech*, vol. 26, no. 9, pp. 1003–1010, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nbt.1487.

[15] Z. Li and C. Chan, "Integrating gene expression and metabolic profiles," *The Journal of Biological Chemistry*, vol. 279, no. 26, pp. 27124–27137, 2004.

[16] P. H. Bradley, M. J. Brauer, J. D. Rabinowitz, and O. G. Troyanskaya, "Coordinated Concentration Changes of Transcripts and Metabolites in Saccharomyces cerevisiae," *PLoS Comput Biol*, vol. 5, no. 1, pp. e1000270+, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1371/journal.pcbi.1000270.

[17] M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson, "Integrating high-throughput and computational data elucidates bacterial networks," *Nature*, vol. 429, no. 6987, pp. 92–96, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nature02456.

[18] C. H. Yeang and M. Vingron, "A joint model of regulatory and metabolic networks," *BMC Bioinformatics*, vol. 7, no. 1, p. 332, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/1471-2105-7-332.

[19] K. Yizhak, W. Benyamini, T.. Liebermeister, E. Ruppin, and T. Shlomi, "Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model," *Bioinformatics (Oxford, England)*, vol. 26, no. 12, pp. i255–260, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/btq183.

[20] A. Manichaikul, L. Ghamsari, E. F. Y. Hom, C. Lin, R. R. Murray, R. L. Chang, S. Balaji, T. Hao, Y. Shen, A. K. Chavali, I. Thiele, X. Yang, C. Fan, E. Mello, D. E. Hill, M. Vidal, K. Salehi-Ashtiani, and J. A. Papin, "Metabolic network analysis integrated with transcript verification for sequenced genomes," *Nature Methods*, vol. 6, no. 8, pp. 589–592, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nmeth.1348.

[21] M. Bylesjö, D. Eriksson, M. Kusano, T. Moritz, and J. Trygg, "Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data," *The Plant Journal*, vol. 52, no. 6, pp. 1181–1191, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1111/j.1365-313X.2007.03293.x.

[22] C. Vijayendran, A. Barsch, K. Friehs, K. Niehaus, A. Becker, and E. Flaschel, "Perceiving molecular evolution processes in Escherichia coli by comprehensive metabolite and gene expression profiling," *Genome Biology*, vol. 9, no. 4, pp. R72+, 2008. Online publication. Available also as printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/gb-2008-9-4-r72.

[23] H. Hotelling, "Relations Between Two Sets of Variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.2307/2333955.

[24] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[25] H. D. Vinod, "Canonical ridge and econometrics of joint production," *Journal of Econometrics*, vol. 4, no. 2, pp. 147–166, 1976.

[26] A. Baccini, P. G. P. Martin, S. Déjean, and I. González, "Cca: An r package to extend canonical correlation analysis," *Journal of Statistical Software*, vol. 23, no. 12, pp. 1–14, undated. Online publication. The journal appears also as printed. Cited 21.9.2011. Available at `http://www.jstatsoft.org/v23/i12`.

[27] K. Lê Cao, I. González, and S. Déjean, "integromics: an r package to unravel relationships between two omics datasets," *Bioinformatics*, vol. 25, no. 21, pp. 2855–2856, 2009. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.1093/bioinformatics/btp515.

[28] D. J. Wilkinson, "Bayesian methods in bioinformatics and computational systems biology," *Briefings in Bioinformatics*, vol. 8, no. 2, pp. 109–116, 2007.

Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bib/bbm007.

[29] F. R. Bach and M. I. Jordan, "A Probabilistic Interpretation of Canonical Correlation Analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley. Online publication. Cited 21.9.2011. Available at `http://stat-ftp.berkeley.edu/tech-reports/688.pdf`.

[30] A. Klami and S. Kaski, "Generative models that discover dependencies between data sets," in *Proceedings of MLSP'06, IEEE International Workshop on Machine Learning for Signal Processing* (S. McLoone, T. Adali, J. Larsen, M. Van Hulle, A. Rogers, and S. Douglas, eds.), pp. 123–128, IEEE, 2006.

[31] A. Klami and S. Kaski, "Probabilistic approach to detecting dependencies between data sets," *Neurocomputing*, vol. 72, pp. 39–46, 2008. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.1016/j.neucom.2007.12.044.

[32] T. Huopaniemi, I. . Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski, "Multivariate multi-way analysis of multi-source data," *Bioinformatics (Oxford, England)*, vol. 26, no. 12, pp. i391–i398, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/btq174.

[33] S. Virtanen, A. Klami, and S. Kaski, "Bayesian cca via group sparsity," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (L. Getoor and T. Scheffer, eds.), ICML '11, (New York, NY, USA), pp. 457–464, ACM, June 2011.

[34] D. J. C. Mackay, "Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 469–505, 1995.

[35] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," vol. 10, no. 3, pp. 515–534, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1093/biostatistics/kxp008.

[36] S. Waaijenborg, P. C. Verselewel de Witt Hamer, and A. H. Zwinderman, "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis," *Statistical applications in genetics and molecular biology*, vol. 7, no. 1, 2008. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.2202/1544-6115.1329.

[37] K. Le Cao, P. Martin, C. Robert-Granie, and P. Besse, "Sparse canonical methods for biological data integration: application to a cross-platform study," *BMC Bioinformatics*, vol. 10, no. 1, p. 34, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1186/1471-2105-10-34.

[38] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, p. Article 1., 2009. Online publication. Cited 21.9.2011. DOI: http://dx.doi.org/10.2202/1544-6115.1406.

[39] P. Mohammadi, "Bayesian integrative modelling of metabolic and transcriptional data using pathway information," Master's thesis, Master of Bioinformatics, Helsinki University of Technology, Espoo, 2010.

[40] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*. New York: W. H. Freeman, 4 ed., 2004.

[41] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, pp. 561–563, 1970.

[42] H. Pearson, "Genetics: What is a gene?," *Nature*, vol. 441, no. 7092, pp. 398–401, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/441398a.

[43] J. T. Kadonaga, "Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors," *Cell*, vol. 116, no. 2, pp. 247–257, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/S0092-8674(03)01078-X.

[44] I. Jordan, L. Marinoramirez, and E. Koonin, "Evolutionary significance of gene expression divergence," *Gene*, vol. 345, no. 1, pp. 119–126, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/j.gene.2004.11.034.

[45] C. Furusawa and K. Kaneko, "Zipf's law in gene expression," *Physical Review Letters*, vol. 90, no. 8, p. 088102, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1103/PhysRevLett.90.088102.

[46] P. Carninci, "Is sequencing enlightenment ending the dark age of the transcriptome?," *Nature Methods*, vol. 6, no. 10, pp. 711–713, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nmeth1009-711.

[47] L. e. a. Sharova, "Database for mrna half-life of 19,977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells," *DNA Research*, vol. 16, p. 45, 2009.

[48] Y. Wang, C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown, "Precision and functional specificity in mRNA decay," *PNAS*, vol. 99, no. 9, pp. 5860–5865, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.092538799.

[49] G. Hambraeus, C. von Wachenfeldt, and L. Hederstedt, "Genome-wide survey of mRNA half-lives in Bacillus subtilis identifies extremely stable mRNAs," *Molecular Genetics and Genomics*, vol. 269, no. 5, pp. 706–714, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1007/s00438-003-0883-6.

[50] D. Gershon, "Dna microarrays: More than gene expression," *Nature*, vol. 437, pp. 1195–1198, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/4371195a.

[51] C. L. Yauk, M. L. Berndt, A. Williams, and G. R. Douglas, "Comprehensive comparison of six microarray technologies," *Nucleic Acids Res*, vol. 32, no. 15, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/nar/gnh123.

[52] J. D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis," *Nature reviews. Genetics*, vol. 7, no. 3, pp. 200–210, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nrg1809.

[53] G. Hardiman, "Microarray platforms–comparisons and contrasts," *Pharmacogenomics*, vol. 5, no. 5, pp. 487–502, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1517/14622416.5.5.487.

[54] T. K. Karakach, S. E. Douglas, and P. D. Wentzell, "An introduction to dna microarrays for gene expression analysis," *Chemometr Intell Lab*, vol. 104, no. 9, pp. 28–52, 2010.

[55] J. Bradford, Y. Hey, T. Yates, Y. Li, S. Pepper, and C. Miller, "A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling," *BMC Genomics*, vol. 11, no. 1, p. 282, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/1471-2164-11-282.

[56] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature biotechnology*, vol. 14, no. 13, pp. 1675–1680, 1996. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nbt1296-1675.

[57] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *Science*, vol. 251, no. 4995, pp. 767–773, 1991. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.1990438.

[58] C. Lausted, T. Dahl, C. Warren, K. King, K. Smith, M. Johnson, R. Saleem, J. Aitchison, L. Hood, and S. R. Lasky, "POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer," *Genome biology*, vol. 5, no. 8, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/gb-2004-5-8-r58.

[59] R. G. Sosnowski, E. Tu, W. F. Butler, J. P. Connell, and M. J. Heller, "Rapid determination of single base mismatch mutations in dna hybrids by direct electric field control," *Proc Natl Acad Sci U S A*, vol. 94, no. 4, pp. 1119–23, 1997.

[60] L. Shi, W. Tong, F. Goodsaid, F. W. Frueh, H. Fang, T. Han, J. C. Fuscoe, and D. A. Casciano, "Qa/qc: challenges and pitfalls facing the microarray community and regulatory agencies," *Expert Rev Mol Diagn*, vol. 4, no. 6, pp. 761–77, 2004.

[61] Editorial, "Making the most of microarrays.," *Nature Biotechnology*, vol. 24, p. 1039, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nbt0906-1039.

[62] L. Shi, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nbt1239.

[63] S. Rosenfeld, "Characteristics of transcriptional activity in nonlinear dynamics of genetic regulatory networks," *Gene Regul Syst Bio.*, vol. 3, p. 159, 2009.

[64] S. Rosenfeld, "Do dna microarrays tell the story of gene expression?," *Gene Regul Syst Bio.*, vol. 4, pp. 61–73, 2010.

[65] B. Lewin, *Genes IX*. Jones & Bartlett Publishers, 2008.

[66] J. Ioannidis, "Why Most Published Research Findings Are False," *PLoS Med*, vol. 2, no. 8, pp. e124+, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1371/journal.pmed.0020124.

[67] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[68] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/1467-9868.00346.

[69] T. I. Lee, "Transcriptional Regulatory Networks in Saccharomyces cerevisiae," *Science*, vol. 298, pp. 799–804, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.1075090.

[70] J. J. Goeman and P. Bühlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics (Oxford, England)*, vol. 23, no. 8, pp. 980–987, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/btm051.

[71] P. Pavlidis, J. Qin, V. Arango, J. J. Mann, and E. Sibille, "Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex," *Neurochemical Research*, vol. 29, no. 6, pp. 1213–1222, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1023/B:NERE.0000023608.29741.45.

[72] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, G. T. R. Pomeroy, S. L., E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.0506580102.

[73] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–13549, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1073/pnas.0506577102.

[74] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/btn577.

[75] S. Draghici, P. Khatri, K. Tarca, A. L. andd Amin, A. Done, C. Voichita, C. andd Georgescu, and R. Romero, "A systems biology approach for pathway level analysis," *Genome Research*, vol. 17, no. 10, pp. 1537–1545, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1101/gr.6202607.

[76] J. Stelling, "Mathematical models in microbial systems biology," *Current Opinion in Microbiology*, vol. 7, no. 5, pp. 513 – 518, 2004.

[77] D. J. Allocco, I. S. Kohane, and A. J. Butte, "Quantifying the relationship between co-expression, co-regulation and gene function," *BMC Bioinformatics*, vol. 5, no. 1, p. 18, 2004. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/ 10.1186/1471-2105-5-18.

[78] D. D. Lewis, "Representation quality in text classification: an introduction and experiment," in *Proceedings of the workshop on Speech and Natural Language*, HLT '90, (Stroudsburg, PA, USA), pp. 288–295, Association for Computational Linguistics, 1990.

[79] D. D. Lewis, "Evaluating text categorization," in *Proceedings of the workshop on Speech and Natural Language*, HLT '91, (Stroudsburg, PA, USA), pp. 312–318, Association for Computational Linguistics, 1991.

[80] M. Zhu, "Recall, precision and average precision," tech. rep., 2004. Online publication. Cited 21.9.2011. Available at `http://sas.uwaterloo.ca/stats_navigation/techreports/04WorkingPapers/2004-09.pdf`.

[81] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2009.

[82] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[83] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, (New York, NY, USA), pp. 11–18, ACM, 2006.

[84] K. Nybakken, S. A. Vokes, T. Y. Lin, A. P. Mcmahon, and N. Perrimon, "A genome-wide rna interference screen in drosophila melanogaster cells for new components of the hh signaling pathway," *Nat. Genet.*, vol. 37, no. 12, pp. 1323–1332, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/ng1682.

[85] N. Bing and I. Hoeschele, "Genetical genomics analysis of a yeast segregant population for transcription network inference," *Genetics*, vol. 170, no. 2, pp. 533–542, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1534/genetics.105.041103.

[86] J. Jaeger, M. Blagov, D. Kosman, K. N. Kozlov, M., E. Myasnikova, S. Surkova, C., M. Samsonova, D. H. Sharp, and J. Reinitz, "Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster," *Genetics*, vol. 167, no. 4, pp. 1721–1737, 2004. Online

publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1534/genetics.104.027334.

[87] A. Marintchev and G. Wagner, "Translation initiation: structures, mechanisms and evolution," *Q Rev Biophys*, vol. 47, pp. 197–284, 2004.

[88] L. D. Kapp and J. R. Lorsch, "The molecular mechanics of eukaryotic translation," *Annual Review of Biochemistry*, vol. 73, no. 1, pp. 657–704, 2004.

[89] D. Iserentant and W. Fiers, "Secondary structure of mrna and efficiency of translation initiation," *Gene*, vol. 9, no. 1-2, pp. 1–12, 1980.

[90] G. D. Stormo, T. D. Schneider, and L. M. Gold, "Characterization of translational initiation sites in e. coli," *Nucleic Acids Research*, vol. 10, no. 9, pp. 2971–2996, 1982.

[91] A. Ciechanover, "Intracellular protein degradation from a vague idea through the lysosome and the ubiquitin-proteasome system and on to human diseases and drug targeting," *Annals of the New York Academy of Sciences*, vol. 1116, no. 1, pp. 1–28, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1196/annals.1402.078.

[92] M. H. Glickman and A. Ciechanover, "The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction," *Physiol. Rev.*, vol. 82, no. 2, pp. 373–428, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1152/physrev.00027.2001.

[93] X. Ang and H. J. Wade, "Scf-mediated protein degradation and cell cycle control," *Oncogene*, vol. 24, no. 7, pp. 2860–2870, 2005.

[94] T. Ravid and M. Hochstrasser, "Diversity of degradation signals in the ubiquitin-proteasome system," *Nature reviews. Molecular cell biology*, vol. 9, no. 9, pp. 679–690, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1038/nrm2468.

[95] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman, "Global analysis of protein expression in yeast," *Nature*, vol. 425, no. 6959, pp. 737–741, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nature02046.

[96] S. Patterson and R. Aebersold, "Proteomics: the first decade and beyond," *Nature Genetics*, vol. 33, pp. 311–323, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/ng1106.

[97] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: a critical review," *Analytical and Bioanalytical Chemistry*, vol. 389, no. 4, pp. 1017–1031–1031, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1007/s00216-007-1486-6.

[98] J. R. Yates, A. Gilchrist, K. E. Howell, and J. J. M. Bergeron, "Proteomics of organelles and large cellular structures," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 9, pp. 702–714, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nrm1711.

[99] H.-C. S. Yen, Q. Xu, D. M. Chou, Z. Zhao, and S. J. Elledge, "Global Protein Stability Profiling in Mammalian Cells," *Science*, vol. 322, no. 5903, pp. 918–923, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.1160489.

[100] B. Futcher, G. Latter, P. Monardo, C. McLaughlin, and J. Garrels, "A sampling of the yeast proteome," *Mol Cell Biol.*, vol. 19, no. 11, pp. 7357–7368, 1999.

[101] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. Derisi, and J. S. Weissman, "Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise," *Nature*, vol. 441, pp. 840–846, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nature04785.

[102] A. A. Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov, and U. Alon, "Dynamic Proteomics of Individual Cancer Cells in Response to a Drug," *Science*, vol. 322, no. 5907, pp. 1511–1516, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.1160165.

[103] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte, "Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation," *Nature Biotechnology*, vol. 25, no. 1, pp. 117–124, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nbt1270.

[104] S. P. Schrimpf, M. Weiss, L. Reiter, C. H. Ahrens, M. Jovanovic, J. Malmström, E. Brunner, S. Mohanty, M. J. Lercher, P. E. Hunziker, R. Aebersold, C. von Mering, , and M. O. Hengartner, "Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes," *PLoS Biol*, vol. 7, no. 3, pp. e1000048+, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1371/journal.pbio.1000048.

[105] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein, "Comparing protein abundance and mRNA expression levels on a genomic scale," *Genome biology*, vol. 4, no. 9, pp. 117+, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/gb-2003-4-9-117.

[106] H. B. Fraser, A. E. Hirsh, G. Giaever, J. Kumm, and M. B. Eisen, "Noise minimization in eukaryotic gene expression," *PLoS Biol*, vol. 2, no. 6, p. e137, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1371%2Fjournal.pbio.0020137.

[107] J. M. Raser and E. K. O'Shea, "Noise in Gene Expression: Origins, Consequences, and Control," *Science*, vol. 309, pp. 2010–2013, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. Available at `http://www.sciencemag.org/cgi/content/abstract/309/5743/2010`.

[108] T. Tuller and E. Kupiec, M. Ruppin, "Determinants of Protein Abundance and Translation Efficiency in S. cerevisiae," *PLoS Comput Biol*, vol. 3, no. 12, pp. e248+, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1371/journal.pcbi.0030248.

[109] de Sousa A. R., L. O. Penalva, E. M. Marcotte, and C. Vogel, "Global signatures of protein and mRNA expression levels," *Molecular bioSystems*, vol. 5, no. 12, pp. 1512–1526, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1039/b908315d.

[110] M. W. Schmidt, A. Houseman, A. R. Ivanov, and D. A. Wolf, "Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe," *Molecular Systems Biology*, vol. 3, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/msb4100117.

[111] B. Lin, J. T. White, W. Lu, T. Xie, A. G. Utleg, X. Yan, E. C. Yi, P. Shannon, I. Khrebtukova, P. H. Lange, D. R. Goodlett, D. Zhou, T. J. Vasicek, and L. Hood, "Evidence for the Presence of Disease-Perturbed Networks in Prostate Cancer Cells by Genomic and Proteomic Analyses: A Systems Approach to Disease," *Cancer Research*, vol. 65, no. 8, pp. 3081–3091, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1158/0008-5472.CAN-04-3218.

[112] R. D. Unwin, D. L. Smith, D. Blinco, C. L. Wilson, C. J. Miller, C. A. Evans, E. Jaworska, S. A. Baldwin, K. Barnes, A. Pierce, E. Spooncer, and A. D. Whetton, "Quantitative proteomics reveals posttranslational control as a regulatory factor in primary hematopoietic stem cells," *Blood*, vol. 107, no. 12, pp. 4687–4694, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1182/blood-2005-12-4995.

[113] T. Ørntoft, T. Thykjaer, F. Waldman, H. Wolf, and J. E. Celis, "Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas," *Mol Cell Proteomics*, vol. 1, no. 1, pp. 37–45, 2002.

[114] Q. Tian, S. Stepaniants, M. Mao, L. Weng, M. Feetham, M. Doyle, E. Yi, H. Dai, V. Thorsson, J. Eng, D. Goodlett, J. Berger, B. Gunter, P. Linseley,

R. Stoughton, R. Aebersold, S. Collins, W. Hanlon, and L. E. Hood, "Integrated genomic and proteomic analyses of gene expression in mammalian cells.," *Mol Cell Proteomics*, vol. 3, no. 10, pp. 960–969, 2004.

[115] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling," *Science*, vol. 324, no. 5924, pp. 218–223, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1126/science.1168978.

[116] K. Baerenfaller, J. Grossmann, M. A. Grobei, R. Hull, M. Hirsch-Hoffmann, S. Yalovsky, P. Zimmermann, U. Grossniklaus, W. Gruissem, and S. Baginsky, "Genome-Scale Proteomics Reveals Arabidopsis thaliana Gene Models and Proteome Dynamics," *Science*, vol. 320, no. 5878, pp. 938–941, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1126/science.1157956.

[117] R. A. Dixon and D. Strack, "Phytochemistry meets genome analysis, and beyond.....," *Phytochemistry*, vol. 62, no. 6, pp. 815 – 816, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/S0031-9422(02)00712-4.

[118] R. J. Bino, R. D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. J. Nikolau, P. Mendes, U. Roessner-Tunali, M. H. Beale, R. N. Trethewey, E. S. Lange, B. M.. Wurtele, and L. W. Sumner, "Potential of metabolomics as a functional genomics tool," *Trends Plant Sci*, vol. 9, no. 9, pp. 418–425, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/j.tplants.2004.07.004.

[119] D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, and E. Huala, "The Arabidopsis Information Resource (TAIR): gene structure and function annotation," *Nucl. Acids Res.*, p. 965, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/nar/gkm965.

[120] M. Stitt, R. Sulpice, and J. Keurentjes, "Metabolic networks: How to identify key components in the regulation of metabolism and growth," *Plant Physiology*, vol. 152, no. 2, pp. 428–444, 2010.

[121] J. E. Lunn, "Gene families and evolution of trehalose metabolism in plants," *Functional Plant Biology*, vol. 34, no. 6, pp. 550–563, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1071/FP06315.

[122] O. Fiehn, "Metabolomics–the link between genotypes and phenotypes," *Plant Mol Biol*, vol. 48, no. 1-2, pp. 155–171, 2002. Online publication. Available

also as a printed journal. Cited 21.9.2011. Available at: `http://view.ncbi.nlm.nih.gov/pubmed/11860207`.

[123] J. Hagel and P. Facchini, "Plant metabolomics: analytical platforms and integration with functional genomics," *Phytochemistry Reviews*, vol. 7, pp. 479–497, 2008. 10.1007/s11101-007-9086-9.

[124] R. D. Hall, "Plant metabolomics: from holistic hope, to hype, to hot topic," *New Phytologist*, vol. 169, pp. 453–468, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1469-8137.2005.01632.x.

[125] K.-M. Oksman-Caldentey and K. Saito, "Integrating genomics and metabolomics for engineering plant metabolic pathways," *Current Opinion in Biotechnology*, vol. 16, no. 2, pp. 174 – 179, 2005.

[126] N. Schauer and A. R. Fernie, "Plant metabolomics: towards biological function and mechanism," *Trends in Plant Science*, vol. 11, no. 10, pp. 508 – 516, 2006.

[127] K. Saito and F. Matsuda, "Metabolomics for functional genomics, systems biology, and biotechnology," *Annual review of plant biology*, vol. 61, no. 1, pp. 463–489, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1146/annurev.arplant.043008.092035.

[128] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, "Systematic functional analysis of the yeast genome," *Trends Biotechnol*, vol. 16, no. 9, pp. 373–378, 1998. Online publication. Available also as a printed journal. Cited 21.9.2011. Available at: `http://view.ncbi.nlm.nih.gov/pubmed/9744112`.

[129] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," *Bioinformatics*, vol. 16, no. 8, pp. 707–726, 2000. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/16.8.707.

[130] J. Nielsen, "It Is All about Metabolic Fluxes," *J. Bacteriol.*, vol. 185, no. 24, pp. 7031–7035, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1128/JB.185.24.7031-7035.2003.

[131] L. J. Sweetlove, D. Fell, and A. R. Fernie, "Getting to grips with the plant metabolic network," *The Biochemical journal*, vol. 409, no. 1, pp. 27–41, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1042/BJ20071115.

[132] J. Förster, I. Famili, P. Fu, B. O. Palsson, and J. Nielsen, "Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network," *Genome Res*, vol. 13, no. 2, pp. 244–253, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1101/gr.234503.

[133] G. Curien, O. Bastien, M. Robert-Genthon, A. Cornish-Bowden, M. L. Cardenas, and R. Dumas, "Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters," *Molecular Systems Biology*, vol. 5, no. 271, 2009. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/msb.2009.29.

[134] L. Blank, L. Kuepfer, and U. Sauer, "Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast," *Genome Biology*, vol. 6, no. 6, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/gb-2005-6-6-r49.

[135] S. Gerdes, R. Edwards, M. Kubal, M. Fonstein, R. Stevens, and A. Osterman, "Essential genes on metabolic maps," *Current Opinion in Biotechnology*, vol. 17, no. 5, pp. 448 – 456, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/j.copbio.2006.08.006.

[136] J. Behre, T. Wilhelm, A. von Kamp, E. Ruppin, and S. Schuster, "Structural robustness of metabolic networks with respect to multiple knockouts," *Journal of Theoretical Biology*, vol. 252, no. 3, pp. 433 – 441, 2008.

[137] G. Spielbauer, L. Margl, L. C. Hannah, W. Romisch, C. Ettenhuber, A. Bacher, A. Gierl, W. Eisenreich, and U. Genschel, "Robustness of central carbohydrate metabolism in developing maize kernels," *Phytochemistry*, vol. 67, no. 14, pp. 1460 – 1475, 2006.

[138] T. C. Williams, L. Miguet, S. K. Masakapalli, N. J. Kruger, L. J. Sweetlove, and R. G. Ratcliffe, "Metabolic network fluxes in heterotrophic Arabidopsis cells: stability of the flux distribution under different oxygenation conditions," *Plant Physiology*, vol. 148, no. 2, pp. 704–18, 2008.

[139] G. N. Stephanopoulos, A. A. Aristidou, and J. Nielsen, *Metabolic Engineering: Principles and Methodologies*. San Diego, California: Academic Press, 1998.

[140] J. Nielsen and S. Oliver, "The next wave in metabolome analysis," *Trends Biotechnol*, vol. 23(11), pp. 544–546, 2005.

[141] L. M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M. C. Walsh, J. A. Berden, K. M. Brindle, D. B. Kell, J. J. Rowland, H. V. Westerhoff, K. van Dam, and S. G. Oliver, "A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations," *Nature biotechnology*, vol. 19, no. 1, pp. 45–50, 2001. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/83496.

[142] W. Weckwerth, M. E. Loureiro, K. Wenzel, and O. Fiehn, "Differential metabolic networks unravel the effects of silent plant phenotypes," *Proc Natl Acad Sci U S A*, vol. 101, no. 20, pp. 7809–7814, 2004. Online

publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.0303415101.

[143] J.-H. S. Hofmeyr and A. Cornish-Bowden, "Regulating the cellular economy of supply and demand," *FEBS Letters*, vol. 476, no. 1–2, pp. 47 – 51, 2000. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/S0014-5793(00)01668-9.

[144] S. Rossell, C. C. van der Weijden, A. Lindenbergh, A. van Tuijl, C. Francke, B. M. Bakker, and H. V. Westerhoff, "Unraveling the complexity of flux regulation: A new method demonstrated for nutrient starvation in Saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences*, vol. 103, no. 7, pp. 2166–2171, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.0509831103.

[145] F. S. Rolleston, "A theoretical background to the use of measured intermediates in the study of the control of intermediary metabolism," *Curr Top Cell Regul*, vol. 5, pp. 47–54, 1972.

[146] H. Kacser and J. A. Burns, "The control of flux," *Symposia of the Society for Experimental Biology*, vol. 27, pp. 65–104, 1973.

[147] H. Kacser and L. Acerenza, "A universal method for achieving increases in metabolite production," *European Journal of Biochemistry*, vol. 216, no. 2, pp. 361–367, 1993. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1432-1033.1993.tb18153.x.

[148] D. A. Fell and S. Thomas, "Physiological control of metabolic flux: the requirement for multisite modulation," *Biochem J*, vol. 311, no. 1, pp. 35–39, 1995.

[149] M. Stitt, "The first will be last and the last will be first: non-regulated enzymes call the tune?," in *Plant Carbohydrate biochemistry* (J. Burrell, M. Bryant, and N. Kruger, eds.), pp. 1–16, BIOS Scientific publisher, Oxford, 1999.

[150] P. Morandini, "Rethinking metabolic control," *Plant Science*, vol. 176, no. 4, pp. 441 – 451, 2009.

[151] M. Stitt and U. Sonnewald, "Regulation of metabolism in transgenic plants," *Annual Review of Plant Physiology and Plant Molecular Biology*, vol. 46, no. 1, pp. 341–368, 1995. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1146/annurev.pp.46.060195.002013.

[152] P. Geigenberger, M. Stitt, and A. R. Fernie, "Metabolic control analysis and regulation of the conversion of sucrose to starch in growing potato tubers," *Plant, Cell & Environment*, vol. 27, no. 6, pp. 655–673, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1111/j.1365-3040.2004.01183.x.

[153] M. Stitt, J. Lunn, and B. Usadel, "Arabidopsis and primary photosynthetic metabolism – more than the icing on the cake," *The Plant Journal*, vol. 61, no. 6, pp. 1067–1091, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1365-313X.2010.04142.x.

[154] J. W. Graham, T. C. Williams, M. Morgan, A. R. Fernie, R. G. Ratcliffe, and L. J. Sweetlove, "Glycolytic enzymes associate dynamically with mitochondria in response to respiratory demand and support substrate channeling," *The Plant cell*, vol. 19, no. 11, pp. 3723–3738, 2007.

[155] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits (Chapman & Hall/CRC Mathematical & Computational Biology)*. Chapman and Hall/CRC, 1 ed., 2006.

[156] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[157] A. Zaslaver, A. E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. G. Surette, and U. Alon, "Just-in-time transcription program in metabolic pathways," *Nature Genetics*, vol. 36, no. 5, pp. 486–491, 2004. Online publication. Cited 13.11.2011. DOI: http://dx.doi.org/10.1038/ng1348.

[158] Y. Gibon, B. Usadel, O. E. Blaesing, B. Kamlage, M. Hoehne, R. Trethewey, and M. Stitt, "Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes," *Genome Biology*, 2006.

[159] J. Kopka, A. Fernie, W. Weckwerth, Y. Gibon, and M. Stitt, "Metabolite profiling in plant biology: platforms and destinations," *Genome Biology*, vol. 5, no. 6, p. Article 109, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1186/gb-2004-5-6-109.

[160] J. Lisec, N. Schauer, J. Kopka, L. Willmitzer, and A. R. Fernie, "Gas chromatography mass spectrometry-based metabolite profiling in plants," *Nature protocols*, vol. 1, no. 1, pp. 387–396, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1038/nprot.2006.59.

[161] R. L. Last, A. D. Jones, and Y. Shachar-Hill, "Towards the plant metabolome and beyond," *Nat Rev Mol Cell Biol*, vol. 8, no. 2, pp. 167–174, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1038/nrm2098.

[162] W. B. Dunn, N. J. Bailey, and H. E. Johnson, "Measuring the metabolome: current analytical technologies," *Analyst*, vol. 130, no. 5, pp. 606–625, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1039/b418288j.

[163] E. Fridman and E. Pichersky, "Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products," *Current Opinion in Plant Biology*, vol. 8, no. 3, pp. 242 – 248, 2005. Physiology andmetabolism.

[164] K. Dettmer and B. D. Hammock, "Metabolomics–a new exciting field within the "omics" sciences," *Environ Health Perspect*, vol. 112, no. 7, 2004.

[165] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 51–78, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1002/mas.20108.

[166] D. Y. Lee, B. P. Bowen, and T. R. Northen, "Mass spectrometry–based metabolomics, analysis of metabolite-protein interactions, and imaging," *BioTechniques*, vol. 49, no. 2, pp. 557–5653, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.2144/000113451.

[167] R. C. De Vos, S. Moco, A. Lommen, J. J. Keurentjes, R. J. Bino, and R. D. Hall, "Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry," *Nat Protoc*, vol. 2, no. 4, pp. 778–791, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nprot.2007.95.

[168] S. Lammert, A. Rockwood, M. Wang, M. Lee, E. Lee, S. Tolley, J. Oliphant, J. Jones, and R. Waite, "Miniature toroidal radio frequency ion trap mass analyzer," *Journal of The American Society for Mass Spectrometry*, vol. 17, pp. 916–922, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/j.jasms.2006.02.009.

[169] K. Akiyama, E. Chikayama, H. Yuasa, Y. Shimada, T. Tohge, K. Shinozaki, S. T. Hirai, M. Y. Y., J. Kikuchi, and K. Saito, "PRIMe: a Web site that assembles tools for metabolomics and transcriptomics," *In silico biology*, vol. 8, no. 3–4, pp. 339–345, 2008.

[170] S. G. Villas-Bôas, J. Højer-Pedersen, M. Å kesson, J. Smedsgaard, and J. Nielsen, "Global metabolite analysis of yeast: evaluation of sample preparation methods," *Yeast*, vol. 22, no. 14, pp. 1155–1169, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1002/yea.1308.

[171] R. Goodacre, D. Broadhurst, A. Smilde, B. Kristal, J. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, and F. Wulfert, "Proposed minimum reporting standards for data analysis in metabolomics," *Metabolomics*, vol. 3, no. 3, pp. 231–241, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1007/s11306-007-0081-3.

[172] C. F. e. a. Taylor, "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project," *Nature Biotechnology*, vol. 26, no. 8, pp. 889–896, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nbt.1411.

[173] H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A. R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R. J. Bino, R. Hall, J. Kopka, G. A. Lane, B. M. Lange, J. R. Liu, P. Mendes, B. J. Nikolau, S. G. Oliver, N. W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L. W. Sumner, T. Wang, S. Walsh, E. S. Wurtele, and D. B. Kell, "A proposed framework for the description of plant metabolomics experiments and their results," *Nature Biotechnology*, vol. 22, no. 12, pp. 1601–1606, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nbt1041.

[174] C. A. Smith, E. J. Want, G. O Maille, R. Abagyan, and G. Siuzdak, "Xcms, processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1021/ac051437y.

[175] M. Katajamaa and M. Orešič, "Data processing for mass spectrometry-based metabolomics," *Journal of Chromatography A*, vol. 1158, no. 1-2, pp. 318–328, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1016/j.chroma.2007.04.021.

[176] Y. Shinbo, Y. Nakamura, M. Altaf-Ul-Amin, H. Asahi, K. Kurokawa, M. Arita, K. Saito, D. Ohta, D. Shibata, and S. Kanaya, "Knapsack: A comprehensive species-metabolite relationship database," in *Plant Metabolomics* (K. Saito, R. A. Dixon, and L. Willmitzer, eds.), vol. 57 of *Biotechnology in Agriculture and Forestry*, pp. 165–181, Springer Berlin Heidelberg, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1007/3-540-29782-0_13.

[177] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. Mcintosh, "A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution lc-ms," *Bioinformatics*, vol. 22, pp. 1902–1909, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/btl276.

[178] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, and J. Selbig, "Metabolite fingerprinting: detecting biological features by independent component analysis," *Bioinformatics*, vol. 20, no. 15, pp. 2447–2454, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/bth270.

[179] A. K. Smilde, M. J. van der Werf, S. Bijlsma, B. J. van der Werff-van der Vat, and R. H. Jellema, "Fusion of mass spectrometry-based metabolomics data," *Anal Chem*, vol. 77, no. 20, pp. 6729–6736, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1021/ac051080y.

[180] M. Steinfath, D. Groth, J. Lisec, and J. Selbig, "Metabolite profile analysis: from raw data to regression and classification," *Physiologia Plantarum*, vol. 132, no. 2, pp. 150–161, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1111/j.1399-3054.2007.01006.x.

[181] J. Trygg and S. Wold, "Orthogonal projections to latent structures (O-PLS)," *Journal of Chemometrics*, vol. 16, no. 3, pp. 119–128, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1002/cem.695.

[182] J. Trygg and S. Wold, "O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter," *J Chemom*, vol. 17, no. 1, pp. 53–64, 2003. Online publication. Available also as printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1002/cem.775.

[183] M. Y. Hirai, M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito, "Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana," *Proc Natl Acad Sci U S A*, vol. 101, no. 27, pp. 10205–10210, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.0403218101.

[184] M. Orešič, C. Clish, E. Davidov, E. Verheij, J. Vogels, L. Havekes, E. Neumann, A. Adourian, S. Naylor, J. van der Greef, and T. Plasterer, "Phenotype characterisation using integrated gene transcript, protein and metabolite profiling," *Applied Bioinformatics*, vol. 3, no. 4, pp. 205–217, 2004.

[185] H. Rischer, M. Orešič, T. Seppänen-Laakso, M. Katajamaa, F. Lammertyn, W. Ardiles-Diaz, M. C. E. Van Montagu, D. Inzé, K.-M. Oksman-Caldentey, and A. Goossens, "Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in Catharanthus roseus cells," *Proceedings of the National Academy of Sciences*, vol. 103, no. 14, pp. 5614–5619, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org//10.1073/pnas.0601027103.

[186] M. Kusano, A. Fukushima, M. Arita, P. Jonsson, T. Moritz, M. Kobayashi, N. Hayashi, T. Tohge, and K. Saito, "Unbiased characterization of genotype-dependent metabolic regulations by metabolomic approach in Arabidopsis thaliana," *BMC Systems Biology*, vol. 1, no. 1, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/1752-0509-1-53.

[187] M. Bylesjö, R. Nilsson, V. Srivastava, A. Grönlund, A. I. Johansson, S. Jansson, J. Karlsson, T. Moritz, G. Wingsle, and J. Trygg, "Integrated Analysis of Transcript, Protein and Metabolite Data To Study Lignin Biosynthesis in Hybrid Aspen," *Journal of Proteome Research*, vol. 8, no. 1, pp. 199–210, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1021/pr800298s.

[188] M. Lehmann, M. Schwarzländer, T. Obata, S. Sirikantaramas, M. Burow, C. E. Olsen, T. Tohge, M. B. L. Fricker, M. D, A. R. Fernie, L. J. Sweetlove, and M. Laxa, "The metabolic response of arabidopsis roots to oxidative stress is distinct from that of heterotrophic cells in culture and highlights a complex relationship between the levels of transcripts, metabolites, and flux," *Molecular Plant*, vol. 2, no. 3, pp. 390–406, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/mp/ssn080.

[189] J. C. Liao, R. Boscolo, Y.-L. L. Yang, L. M. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: reconstruction of regulatory signals in biological systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 26, pp. 15522–15527, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.2136632100.

[190] A. Fernie, R. Trethewey, A. Krotzky, and L. Willmitzer, "Metabolite profiling: from diagnostics to systems biology," *Nat Rev Mol Cell Biol.*, vol. 5, no. 9, pp. 763–769, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nrm1451.

[191] K. Janes and M. B. Yaffe, "Data-driven modelling of signal-transduction networks," *Nat Rev Mol Cell Biol.*, vol. 7, no. 11, pp. 820–828, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1038/nrm2041.

[192] C. Guy, J. Kopka, , and T. Moritz, "Plant metabolomics coming of age," *Physiologia Plantarum*, vol. 132, no. 2, pp. 113–116, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1399-3054.2007.01020.x.

[193] U. Roessner, A. Luedemann, D. Brust, O. Fiehn, T. Linke, L. Willmitzer, and A. R. Fernie, "Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems," *The Plant Cell Online*, vol. 13, no. 1, pp. 11–29, 2001. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.13.1.11.

[194] A. M. Martins, D. Camacho, J. Shuman, W. Sha, P. Mendes, and V. Shulaev, "A systems biology study of two distinct growth phases of saccharomyces cerevisiae cultures," *Current Genomics*, vol. 5, pp. 649–663, 2004.

[195] K. Yonekura-Sakakibara, T. Tohge, F. Matsuda, R. Nakabayashi, H. Takayama, , R. Niida, A. Watanabe-Takahashi, E. Inoue, and K. Saito, "Comprehensive Flavonol Profiling and Transcriptome Coexpression Analysis Leading to Decoding Gene-Metabolite Correlations in Arabidopsis," *Plant Cell*, pp. tpc.108.058040+, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.108.058040.

[196] R. Steuer, "Review: On the analysis and interpretation of correlations in metabolomic data," *Briefings in Bioinformatics*, vol. 7, no. 2, pp. 151–158, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bib/bbl009.

[197] D. Camacho, A. de la Fuente, and P. Mendes, "The origin of correlations in metabolomics data," *Metabolomics*, vol. 1, no. 1, pp. 53–63, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1007/s11306-005-1107-3.

[198] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth, "Observing and interpreting correlations in metabolomic networks," *Bioinformatics*, vol. 19, no. 8, pp. 1019–1026, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. Available at: `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/8/1019`.

[199] K. Morgenthal, W. Weckwerth, and R. Steuer, "Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation," vol. 83, no. 2-3, pp. 108–117, 2006.

[200] F. Kose, J. Budzies, M. Holschneider, and O. Fiehn, "Robust detection and verification of linear relationships to generate metabolic networks using estimates of technical errors," *BMC Bioinformatics*, vol. 8, p. 162, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/1471-2105-8-162.

[201] R. Steuer, T. Gross, J. Selbig, and B. Blasius, "Structural kinetic modeling of metabolic networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 32, pp. 11868–11873, 2006. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.0600013103.

[202] M. Rantalainen, O. Cloarec, T. Ebbels, T. Lundstedt, J. Nicholson, E. Holmes, and J. Trygg, "Piecewise multivariate modelling of sequential metabolic profiling data," *BMC Bioinformatics*, vol. 9, no. 1, pp. 105+, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/1471-2105-9-105.

[203] K. Urano, K. Maruyama, Y. Ogata, Y. Morishita, M. Takeda, N. Sakurai, H. Suzuki, K. Saito, D. Shibata, K. Kobayashi, M. Yamaguchi-Shinozaki, and

K. Shinozaki, "Characterization of the ABA-regulated global responses to dehydration in Arabidopsis by metabolomics," *The Plant Journal*, vol. 57, no. 6, pp. 1065–1078, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1365-313X.2008.03748.x.

[204] D. Kley, M. Kleinmann, H. Sanderman, and S. Krupa, "Photochemical oxidants: state of the science," *Environmental Pollution*, vol. 100, no. 1–3, pp. 19 –42, 1999. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/S0269-7491(99)00086-X.

[205] H. Wohlgemuth, K. Mittelstrass, S. Kschieschan, J. Bender, H.-J. Weigel, K. Overmyer, J. Kangasjärvi, H. Sandermann, and C. Langebartels, "Activation of an oxidative burst is a general feature of sensitive plants exposed to the air pollutant ozone," *Plant, Cell & Environment*, vol. 25, no. 6, pp. 717–726, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1046/j.1365-3040.2002.00859.x.

[206] M. U. J. Kangasjärvi, J. Talvinen and R. Karjalainen, "Plant defence systems induced by ozone," *Plant, Cell & Environment*, vol. 17, pp. 784–794, 1994.

[207] R. L. Heath and G. E. Taylor, "Physiological processes and plant responses to ozone exposure," in *Forest Decline and Ozone* (H. Sandermann, A. R. Wellburn, and R. L. Heath, eds.), vol. 127 of *Ecological Studies*, pp. 317–368, Springer Berlin Heidelberg, 1997.

[208] E. J. Pell, C. D. Schlagnhaufer, and R. N. Arteca, "Ozone-induced oxidative stress: Mechanisms of action and reaction," *Physiologia Plantarum*, vol. 100, no. 2, pp. 264–273, 1997. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1399-3054.1997.tb04782.x.

[209] C. Langebarterls and J. Kangasjärvi, "Ethylene and jasmonate as regulators or cell death in disease resistance,," in *Molecular Ecotixicology of Plants* (H. Sandermann, ed.), (Berlin, Heidelberg), pp. 75–109, Springer-Verlag, 2004.

[210] Y. G. Sharma and K. R. Davis, "The effects of ozone on antioxidant responses in plants," *Free Radical Biology and Medicine*, vol. 23, no. 3, pp. 480 – 488, 1997.

[211] H. Sandermann Jr, D. Ernst, W. Heller, and C. Langebartels, "Ozone: An abiotic elicitor of plant defence reactions," *Trends in Plant Science*, vol. 3, no. 2, pp. 47 – 50, 1998.

[212] R. A. Dietrich, T. P. Delaney, S. J. Uknes, J. A. Ward, E. R.a nd Ryals, and J. L. Dangl, "Arabidopsis mutants simulating disease resistance response," *Cell*, vol. 77, no. 4, pp. 565–577, 1994. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/0092-8674(94)90218-6.

[213] J. Kangasjärvi, P. Jaspers, and H. Kollist, "Signalling and cell death in ozone-exposed plants," *Plant, Cell & Environment*, vol. 28, no. 8, pp. 1021–1036, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1365-3040.2005.01325.x.

[214] M. Luwe, U. Takahama, and U. Heber, "Role of ascorbate in detoxifying ozone in the apoplast of spinach (spinacia oleracea l.) leaves.," *Plant Physiology*, vol. 101, no. 3, pp. 969–976, 1993.

[215] C. Lamb and R. A. A. Dixon, "The oxidative burst in plant disease resistance," *Annu Rev Plant Physiol Plant Mol Biol*, vol. 48, pp. 251–275, 1997. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/0.1146/annurev.arplant.48.1.251.

[216] M. Baier, A. Kandlbinder, D. Golldack, and K. Dietz, "Oxidative stress and ozone: perception, signalling and response," *Plant, Cell & Environment*, vol. 28, no. 8, pp. 1012–1020, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1365-3040.2005.01326.x.

[217] F. Van Breusegem, E. Vranová, J. F. Dat, and D. Inzé, "The role of active oxygen species in plant signal transduction," *Plant Science*, vol. 161, no. 3, pp. 405 – 414, 2001.

[218] A. N. Dodd, J. Kudla, and D. Sanders, "The language of calcium signaling," *Annual review of plant biology*, vol. 61, pp. 593–620, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1146/annurev-arplant-070109-104628.

[219] P. Jaspers and J. Kangasjärvi, "Reactive oxygen species in abiotic stress signaling," *Physiol Plant*, vol. 138, no. 4, pp. 405–13, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1399-3054.2009.01321.x.

[220] M. C. S. Rodriguez, M. Petersen, and J. Mundy, "Mitogen-activated protein kinase signaling in plants," *Annual Review of Plant Biology*, vol. 61, pp. 621–649, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1146/annurev-arplant-042809-112252.

[221] J. H. Joo, S. Wang, J. Chen, A. Jones, and N. V. Fedoroff, "Different signaling and cell death roles of heterotrimeric g protein α and β subunits in the arabidopsis oxidative stress response to ozone," *The Plant Cell Online*, vol. 17, no. 3, pp. 957–970, 2005. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.104.029603.

[222] R. Mahalingam, N. Shah, A. Scrymgeour, and N. Fedoroff, "Temporal evolution of the *Arabidopsis* oxidative stress response," *Plant Molecular Biology*, vol. 57, pp. 709–730, 2005. 10.1007/s11103-005-2860-4.

[223] B. L. Örvar, J. McPherson, and B. E. Ellis, "Pre-activating wounding response in tobacco prior to high-level ozone exposure prevents necrotic injury," *The Plant Journal*, vol. 11, no. 2, pp. 203–212, 1997. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1046/j.1365-313X.1997.11020203.x.

[224] K. Overmyer, M. Brosché, R. Pellinen, T. Kuittinen, H. Tuominen, R. Ahlfors, . S. M. Keinänen, M, D. Scheel, and J. Kangasjärvi, "Ozone-Induced Programmed Cell Death in the Arabidopsis radical-induced cell death1 Mutant," *Plant Physiol*, 2005.

[225] E. P. Beers and J. M. McDowell, "Regulation and execution of programmed cell death in response to pathogens, stress and developmental cues," *Current Opinion in Plant Biology*, vol. 4, no. 6, pp. 561 – 567, 2001.

[226] M. V. Rao and K. R. Davis, "The physiology of ozone induced cell death," *Planta*, vol. 213, pp. 682–690, 2001. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1007/s004250100618.

[227] S. Berger, "Jasmonate-related mutants of arabidopsis as tools for studying stress signaling," *Planta*, vol. 214, pp. 497–504, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1007/s00425-001-0688-y.

[228] J. T. Greenberg and F. M. Ausubel, "Arabidopsis mutants compromised for the control of cellular damage during pathogenesis and aging," *The Plant Journal*, vol. 4, no. 2, pp. 327–341, 1993. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1046/j.1365-313X.1993.04020327.x.

[229] J. T. Greenberg, A. Guo, D. F. Klessig, and F. M. Ausubel, "Programmed cell death in plants: A pathogen-triggered response activated coordinately with multiple defense functions," *Cell*, vol. 77, no. 4, pp. 551–563, 1994. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1016/0092-8674(94)90217-8.

[230] D. N. Rate, J. V. Cuenca, G. R. Bowman, D. S. Guttman, and J. T. Greenberg, "The gain-of-function arabidopsis acd6 mutant reveals novel regulation and function of the salicylic acid signaling pathway in controlling cell death, defenses, and cell growth.," *Plant Cell*, vol. 11, no. 9, pp. 1695–708, 1999.

[231] K. Overmyer, H. Tuominen, R. Kettunen, C. Betz, C. Langebartels, H. Sandermann, and J. Kangasjärvi, "Ozone-sensitive arabidopsis rcd1 mutant reveals opposite roles for ethylene and jasmonate signaling pathways in regulating superoxide-dependent cell death," *The Plant Cell Online*, vol. 12, no. 10, pp. 1849–1862, 2000. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.12.10.1849.

[232] K. Overmyer, M. Broscheé, and Kangasj "Reactive oxygen species and hormonal control of cell death," *Trends in Plant Science*, vol. 8, no. 7, pp. 335 – 342, 2003.

[233] J. Tuomainen, C. Betz, J. Kangasjärvi, D. Ernst, Z. Yin, C. Langebartels, and H. Sandermann, "Ozone induction of ethylene emission in tomato plants: regulation by differential accumulation of transcripts for the biosynthetic enzymes," *The Plant Journal*, vol. 12, no. 5, pp. 1151–1162, 1997. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1046/j.1365-313X.1997.12051151.x.

[234] M. V. Rao, J. R. Koch, and K. R. Davis, "Ozone: a tool for probing programmed cell death in plants," *Plant Molecular Biology*, vol. 44, pp. 345–358, 2000. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1023/A:1026548726807.

[235] M. V. Rao, H. Lee, R. A. Creelman, J. E. Mullet, and K. R. Davis, "Jasmonic acid signaling modulates ozone-induced hypersensitive cell death," *The Plant Cell Online*, vol. 12, no. 9, pp. 1633—1646, 2000. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.12.9.1633.

[236] W. Moeder, C. S. Barry, C. Tauriainen, A. A. andBetz, J. Tuomainen, M. Utriainen, D. Grierson, H. Sandermann, C. Langebartels, and J. Kangasjärvi, "Ethylene synthesis regulated by biphasic induction of 1-aminocyclopropane-1-carboxylic acid synthase and 1-aminocyclopropane-1-carboxylic acid oxidase genes is required for hydrogen peroxide accumulation and cell death in ozone-exposed tomato," *Plant Physiology*, vol. 130, no. 4, pp. 1918–1926, 2002.

[237] M. Kanna, M. Tamaoki, A. Kubo, N. Nakajima, R. Rakwal, G. K. Agrawal, S. Tamogami, M. Ioki, D. Ogawa, H. Saji, and M. Aono, "Isolation of an ozone-sensitive and jasmonate-semi-insensitive arabidopsis mutant (oji1)," *Plant and Cell Physiology*, vol. 44, no. 12, pp. 1301–1310, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/pcp/pcg157.

[238] H. Tuominen, K. Overmyer, M. Keinänen, H. Kollist, and J. Kangasjärvi, "Mutual antagonism of ethylene and jasmonic acid regulates ozone-induced spreading cell death in arabidopsis," *The Plant Journal*, vol. 39, no. 1, pp. 59–69, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1365-313X.2004.02107.x.

[239] M. V. Rao and K. R. Davis, "Ozone-induced cell death occurs via two distinct mechanisms in arabidopsis: the role of salicylic acid," *The Plant Journal*, vol. 17, no. 6, pp. 603–614, 1999. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1046/j.1365-313X.1999.00400.x.

[240] W. Van Camp, M. Van Montagu, and D. Inzé, "H_2o_2 and no: redox signals in disease resistance," *Trends in Plant Science*, vol. 3, pp. 330–334, 1998.

[241] A. M. Mazzucotelli, E. Mastrangelo, C. Crosatti, D. Guerra, A. M. Stanca, and L. Cattivelli, "Abiotic stress response in plants: When post-transcriptional and post-translational regulations control transcription," *Plant Science*, vol. 174, no. 4, pp. 420 – 431, 2008.

[242] K. Shinozaki, K. Y. Shinozaki, and M. Seki, "Regulatory network of gene expression in the drought and cold stress responses," *Curr Opin Plant Biol*, vol. 6, no. 5, pp. 410–417, 2003.

[243] J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. Bornberg-Bauer, J. Kudla, and K. Harter, "The atgenexpress global stress expression data set: protocols, evaluation and model data analysis of uv-b light, drought and cold stress responses," *The Plant Journal*, vol. 50, no. 2, pp. 347–363, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1111/j.1365-313X.2007.03052.x.

[244] N. Mitsuda and M. Ohme-Takagi, "Functional analysis of transcription factors in arabidopsis," *Plant and Cell Physiology*, vol. 50, no. 7, pp. 1232–1248, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/pcp/pcp075.

[245] I. Gadjev, S. Vanderauwera, T. S. Gechev, C. Laloi, I. N. Minkov, V. Shulaev, K. Apel, D. Inzé, R. Mittler, and F. Van Breusegem, "Transcriptomic footprints disclose specificity of reactive oxygen species signaling in arabidopsis," *Plant Physiology*, vol. 141, no. 2, pp. 436–445, 2006.

[246] L. Vaahtera and M. Brosché, "More than the sum of its parts - how to achieve a specific transcriptional response to abiotic stress," *Plant Science*, vol. 180, no. 3, pp. 421 – 430, 2011.

[247] K. M. Herrmann, "The shikimate pathway: Early steps in the biosynthesis of aromatic compounds," *The Plant Cell Online*, vol. 7, no. 7, pp. 907–919, 1995. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.7.7.907.

[248] A. W. Woodward and B. Bartel, "Auxin: Regulation, action, and interaction," *Annals of Botany*, vol. 95, no. 5, pp. 707–735, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/aob/mci083.

[249] A. Ishihara, F. Matsuda, H. Miyagawa, and K. Wakasa, "Metabolomics for metabolically manipulated plants: effects of tryptophan overproduction," *Metabolomics*, vol. 3, pp. 319–334, 2007. 10.1007/s11306-007-0072-4.

[250] R. Ahlfors, S. Lång, K. Overmyer, P. Jaspers, M. Broscheé, A. Tauriainen, H. Kollist, H. Tuominen, E. Belles-Boix, M. Piippo, D. Inzé, E. T. Palva, and J. Kangasjärvi, "Arabidopsis radical-induced cell death1 belongs to the wwe protein–protein interaction domain protein family and modulates abscisic acid, ethylene, and methyl jasmonate responses," *The Plant Cell Online*, vol. 16, no. 7, pp. 1925–1937, 2004. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.021832.

[251] T. Fujibe, H. Saji, K. Arakawa, N. Yabe, Y. Takeuchi, and K. Yamamoto, "A methyl viologen-resistant mutant of arabidopsis, which is allelic to ozone-sensitive rcd1, is tolerant to supplemental ultraviolet-b irradiation," *Plant Physiology*, vol. 134, no. 1, pp. 275–85, 2004.

[252] E. Belles-Boix, V. M. M. Babiychuk, E., D. Inzé, and S. Kushnir, "Ceo1, a new protein from arabidopsis thaliana, protects yeast against oxidative damage," *FEBS Letters*, vol. 482, no. 1-2, pp. 19 – 24, 2000.

[253] Q. Liu, M. Kasuga, Y. Sakuma, H. Abe, S. Miura, K. Yamaguchi-Shinozaki, and K. Shinozaki, "Two transcription factors, dreb1 and dreb2, with an erebp/ap2 dna binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in arabidopsis," *The Plant Cell Online*, vol. 10, no. 8, pp. 1391–1406, 1998. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1105/tpc.10.8.1391.

[254] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing ed., 2007.

[255] I. M. Johnstone and D. M. Titterington, "Statistical challenges of high-dimensional data," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 367, no. 1906, pp. 4237–4253, 2009.

[256] A. Buja, T. Hastie, and R. Tibshirani, "Linear smoothers and additive models," *The Annals of Statistics*, vol. 17, pp. 453–510, 1989.

[257] I. Morlini, "On multicollinearity and concurvity in some nonlinear multivariate models," *Statistical Methods and Applications*, vol. 15, no. 1, pp. 3–26, 2006.

[258] P. Bickel, B. Li, A. Tsybakov, S. van de Geer, B. Yu, , T. Valdés, C. Rivero, J. Fan, and A. van der Vaart, "Regularization in statistics," *TEST*, vol. 15, no. 2, pp. 271–344, 2006. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.1007/BF02607055.

[259] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000.

[260] A. Gelman, C., J. B., H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 2 ed., 2003.

[261] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1023/A:1007665907178.

[262] T. Jaakkola, "Tutorial on variational approximation methods," in *Advanced mean field methods: theory and practice* (M. Opper and D. Saad, eds.), MIT Press, 2000.

[263] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[264] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[265] F. Galton, "Regression Towards Mediocrity in Hereditary Stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 15, pp. 246–263, 1886. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.2307/2841583.

[266] K. Pearson, "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.2307/90707.

[267] M. Borga, "Canonical correlation a tutorial," 1999. Online publication. Updated 29.10.1999. Cited 21.9.2011. Available at: `http://www.imt.liu.se/~magnus/cca/tutorial/`.

[268] H. Hotelling, "Analysis of complex statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[269] M. S. Bartlett, "The statistical significance of cononical correlations," *Biometrika*, vol. 32, no. 1, pp. 29–37, 1941.

[270] Y. Takane, H. Yanai, and H. Hwang, "An improved method for generalized constrained canonical correlation analysis," *Computational Statistics & Data Analysis*, vol. 50, no. 1, pp. 221–241, 2006. Online publication. The journal appears also as printed. Cited 21.9.2011. Available at: `http://ideas.repec.org/a/eee/csdana/v50y2006i1p221-241.html`.

[271] T. Rinnet, "Extracting stress-related effects from yeast gene expression by canonical correlation analysis," Master's thesis, Technical Physics, Helsinki University of Technology, Espoo, 2006.

[272] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis – a global perspective*. Upper Saddle River, New Jersey: Pearson Education, 7 ed., 2010.

[273] W. Meredith, "Canonical correlations with fallible data," *Psychometrika*, vol. 29, no. 1, pp. 55–65, 1964.

[274] C. J. F. Ter Braak, "Interpreting canonical correlation analysis through biplots of structural correlations and weights," *Psychometrika*, vol. 55, pp. 519–531, 1990.

[275] R. M. Thorndike and D. J. Weiss, "A Study of the Stability of Canonical Correlations and Canonical Components," vol. 33, no. 1, pp. 123–134, 1973. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1177/001316447303300113.

[276] B. Thompson, "Exploring the Replicability of a Study's Results: Bootstrap Statistics for the Multivariate Case," *Educational and Psychological Measurement*, vol. 55, no. 1, pp. 84–94, 1995. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1177/0013164495055001008.

[277] B. Thompson, *Canonical correlation analysis: uses and interpretation.* Beverly Hills, California: Sage Publications Inc, 1984.

[278] K. Nimon, R. K. Henson, and M. S. Gates, "Revisiting interpretation of canonical correlation analysis: A tutorial and demonstration of canonical commonality analysis," *Multivariate Behavioral Research*, vol. 45, no. 4, pp. 702–724, 2010.

[279] D. Stewart and W. Love, "A general canonical correlation index," *Psychol Bull.*, vol. 70, no. 3, pp. 160–163, 1968.

[280] J. Caldas, F. A. B. A. Gehlenborg, N., and S. Kaski, "Probabilistic retrieval and visualization of biologically relevant microarray experiments," *Bioinformatics*, vol. 25, no. 12, pp. i145–i153, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bioinformatics/btp215.

[281] M. L. Eaton and M. D. Perlman, "The Non-Singularity of Generalized Sample Covariance Matrices," *The Annals of Statistics*, vol. 1, no. 4, pp. 710–717, 1973. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1214/aos/1176342465.

[282] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, "Canonical Correlation Analysis when the Data are Curves," *Journal of the Royal Statistical Society. Series B*, no. 3, pp. 725–740.

[283] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, (San Francisco, CA, USA), pp. 1137–1145, Morgan Kaufmann Publishers Inc., 1995.

[284] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.2307/2958830.

[285] S. Waaijenborg and A. Zwinderman, "Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks," *BMC Bioinformatics*, vol. 10, no. 1, pp. 315+, 2009. Online publication. Available also as printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/1471-2105-10-315.

[286] S. Waaijenborg and A. H. Zwinderman, "Penalized canonical correlation analysis to quantify the association between gene expression and dna markers," *BMC Proceedings*, vol. 1, no. Suppl 1, p. S122, 2007. Online publication. Cited 21.9.2011. Available at: `http://www.biomedcentral.com/1753-6561/1/S1/S122`.

[287] I. González, S. Déjean, P. Martin, O. Gon¸calves, P. Besse, and A. Baccini, "Highlighting relationships between heterogenous biological data through graphical displays based on regularized canonical correlation analysis," *Journal of Biological Systems*, vol. 17, no. 2, pp. 173–199, 2009.

[288] A. Baccini, P. Besse, S. Déjean, P. Martin, C. Robert-Granié, and S. C. M., "Stratégies pour l'analyse statistique de donné es transcriptomiques," *Journal de la Société Française de Statistique*, vol. 146, pp. 4–44, 2005.

[289] P. G. P. Martin, H. Guillou, F. Lasserre, S. D éjean, A. Lan, J. Pascussi, M. SanCristobal, P. Legrand, P. Besse, and T. Pineau, "Novel aspects of pparα-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study," *Hepatology*, vol. 45, no. 3, pp. 767–777, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1002/hep.21510.

[290] S. Jozefczuk, S. Klie, G. Catchpole, J. Szymanski, A. Cuadros-Inostroza, D. Steinhauser, J. Selbig, and L. Willmitzer, "Metabolomic and transcriptomic stress response of Escherichia coli," *Molecular Systems Biology*, vol. 6, p. 364, 2010.

[291] C. Archambeau, N. Delannay, and M. Verleysen, "Robust probabilistic projections," in *ICML '06 Proceedings of the 23rd international conference on Machine learning*, ICML '06, (New York, NY, USA), pp. 33–40, ACM, 2006.

[292] M. E. Tipping and C. M. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, pp. 611–622, 1999. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.1111/1467-9868.00196.

[293] A. Klami, S. Virtanen, and S. Kaski, "Bayesian exponential family projections for coupled data sources," in *Proceedings of the Twenty-Sixth Conference on*

Uncertainty in Artificial Intelligence (2010) (P. Grunwald and P. Spirtes, eds.), (Corvallis, Oregon), pp. 286–293, AUAI Press, 2010.

[294] C. Archambeau and F. Bach, "Sparse probabilistic projections," in *Advances in Neural Information Processing Systems 21* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 17–24, The MIT Press, 2009.

[295] C. M. Bishop, "Bayesian pca," in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, (Cambridge, MA, USA), pp. 382–388, MIT Press, 1999.

[296] A. Klami and S. Kaski, "Local dependent components," in *Proceedings of the 24th international conference on Machine learning* (Z. Ghahramani, ed.), ICML '07, (New York, NY, USA), pp. 425–432, ACM, 2007.

[297] P. Rai and H. Daumé, "Multi-label prediction via sparse infinite cca," in *Proceedings of the Conference on Neural Information Processing Systems 22* (S. D. L. J. W. C. K. I. Bengio, Y. and A. Culotta, eds.), (Vancouver, Canada), pp. 1518–1526, 2009.

[298] C. Wang, "Variational bayesian approach to canonical correlation analysis," *IEEE Transactions on Neural Networks*, vol. 18, no. 3, pp. 905–910, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1109/TNN.2007.891186.

[299] Z. Ghahramani, T. L. Griffiths, and P. Sollich, *Bayesian nonparametric latent feature models*. Oxford: Oxford University Press, 2007.

[300] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Advances in Neural Information Processing Systems 23* (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, eds.), pp. 982–990, 2010.

[301] M. Salzmann, C. H. Ek, R. Urtasun, and T. Darrell, "Factorized Orthogonal Latent Spaces," *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 701–708, 2010.

[302] C. M. Bishop, "Variational principal components," in *In Proceedings Ninth International Conference on Artificial Neural Networks, ICANN '99*, pp. 509–514, 1999. Online publication. Cited 21.9.2011. DOI: http://dx.doi.org/10.1049/cp:19991160.

[303] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, "Estimating image bases for visual image reconstruction from human brain activity," in *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 576–584, 2009.

[304] J. Luttinen and A. Ilin, "Transformations in variational bayesian factor analysis to speed up learning," *Neurocomput.*, vol. 73, no. 7–9, pp. 1093–1102, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1016/j.neucom.2009.11.018.

[305] K. F. Kerr, "Extended analysis of benchmark datasets for Agilent two-color microarrays," *BMC Bioinformatics*, vol. 8, no. 1, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1186/1471-2105-8-371.

[306] Brechbuehler, "Amino acid analysis," 2011. Online publication. Cited 21.7.2011. Available at: `http://www.brechbuehler.ch/EZ-faast.191.0.html`.

[307] L. A. Mueller, P. Zhang, and S. Y. Rhee, "AraCyc: A Biochemical Pathway Database for Arabidopsis," *Plant Physiology*, vol. 132, no. 2, pp. 453–460, 2003.

[308] P. Zhang, H. Foerster, C. P. Tissier, L. Mueller, S. Paley, P. D. Karp, and S. Y. Rhee, "Metacyc and aracyc. metabolic pathway databases for plant research," *Plant Physiology*, vol. 138, no. 1, pp. 27–37, 2005.

[309] M. G. Poolman, L. Miguet, L. J. Sweetlove, and D. A. Fell, "A Genome-Scale Metabolic Model of Arabidopsis and Some of Its Properties," *Plant Physiol.*, vol. 151, no. 3, pp. 1570–1581, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1104/pp.109.141267.

[310] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi, "Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology," *Briefings in Bioinformatics*, vol. 11, no. 1, pp. 40–79, 2010. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1093/bib/bbp043.

[311] H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis* (P. R. Krishnaiah, ed.), pp. 391–420, 1966.

[312] J. Thioulouse, D. Chessel, S. Dolédec, and J. Olivier, "Ade-4: a multivariate analysis and graphical display software," *Stat. Comput*, vol. 7, pp. 75–83, 1997.

[313] J. C. Pinheiro and D. M. Bates, *Mixed Effects Models in S and S-Plus*. New York: Springer, 1 ed., 2000.

[314] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Texts in Statistics, New York: Springer, 3 ed., 2005.

[315] J. C. Pinheiro, B. M. Douglas, S. DebRoy, D. Sarkar, and R Development Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2011. R package version 3.1-100.

[316] M. F. Covington and S. L. Harmer, "The circadian clock regulates auxin signaling and responses in *Arabidopsis*," *PLoS Biol*, vol. 5, no. 8, p. e222, 2007. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1371%2Fjournal.pbio.0050222.

[317] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1023/A:1012487302797.

[318] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/.

[319] K. A. Lê Cao, O. Gonçalves, P. Besse, and S. Gadat, "Selection of biologically relevant genes with a wrapper stochastic algorithm," *Statistical applications in genetics and molecular biology*, vol. 6, p. Article 29, 2007. Online publication. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.2202/1544-6115.1312.

[320] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996. Online publication. The journal appears also as printed. Cited 21.9.2011. Available at: `http://www.ams.org/mathscinet-getitem?mr=1379242`.

[321] K. A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse PLS for variable selection when integrating omics data," *Statistical applications in genetics and molecular biology*, vol. 7, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.2202/1544-6115.1390.

[322] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 2, pp. 301–320, 2005. Online publication. The journal appears also as printed. Cited 21.9.2011. DOI: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

[323] S. Waaijenborg and A. H. Zwinderman, "Correlating multiple SNPs and multiple disease phenotypes: penalized non-linear canonical correlation analysis," *Bioinformatics*, vol. 25, no. 21, pp. 2764–2771, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1093/bioinformatics/btp491.

[324] E. Parkhomenko, D. Tritchler, and J. Beyene, "Genome-wide sparse canonical correlation of gene expression with genotypes," *BMC Proceedings*, vol. 1, no. Suppl 1, p. S119, 2007. Online publication. Available also as a printed

journal. Cited 21.9.2011. Available at: http://www.biomedcentral.com/1753-6561/1/S1/S119.

[325] T. Park and G. Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: http://dx.doi.org/10.1198/016214508000000337.

[326] Q. Li and N. Lin, "The bayesian elastic net," *Bayesian Analysis*, vol. 5, no. 1, pp. 151–170, 2010.

[327] J. de Leeuw, Y. F., and Y. Takane, "Additive structure in qualitative data: An alternating least squares method with optimal scaling features," *Psychometrika*, vol. 41, no. 4, pp. 471–503, 1976.

[328] E. van der Burg and J. de Leeuw, "Non-linear canonical correlation," *British Journal of Mathematical and Statistical Psychology*, vol. 36, pp. 54–80, 1983.

[329] E. van der Burg and J. de Leeuw, "Nonlinear canonical correlation analysis with k sets of variables," *Tech rep 87-7*, 1987.

[330] E. van der Burg, J. de Leeuw, and G. Dijksterhuis, "Nonlinear canonical correlation with k sets of variables," *Computational Statistics & Data Analysis*, vol. 18, no. 1, pp. 141–163, 1994.

[331] F. Young, J. de Leeuw, and Y. Takane, "Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features," *Psychometrika*, vol. 41, no. 4, pp. 505–529, 1976.

[332] J.-G. G. Joung, A. M. Corbett, S. Fellman, M. Moore, D. M. Tieman, H. J. Klee, J. J. Giovannoni, and Z. Fei, "Plant MetGenMAP: an integrative analysis system for plant systems biology," *Plant physiology*, vol. 151, no. 4, pp. 1758–1768, 2009. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1104/pp.109.145169.

[333] O. Thimm, O. Blasing, Y. Gibon, A. Nagel, S. Meyer, P. Kruger, J. Selbig, L. A. Muller, S. Y. Rhee, and M. Stitt, "MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes," vol. 37, no. 6, pp. 914–939, 2004. Online publication. Available also as a printed journal. Cited 21.9.2011. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14996223.

[334] T. Tokimatsu, N. Sakurai, H. Suzuki, H. Ohta, K. Nishitani, T. Koyama, T. Umezawa, N. Misawa, K. Saito, and D. Shibata, "Kappa-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps," *Plant Physiol*, vol. 138, no. 3, pp. 1289–300, 2005.

[335] K. R. Patil and J. Nielsen, "Uncovering transcriptional regulation of metabolism by using metabolic network topology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 8, pp. 2685–2689, 2005. Online publication. Available also as a printed journal. Cited 21.9.2011. DOI: htpp://dx.doi.org/10.1073/pnas.0406811102.

Appendix A

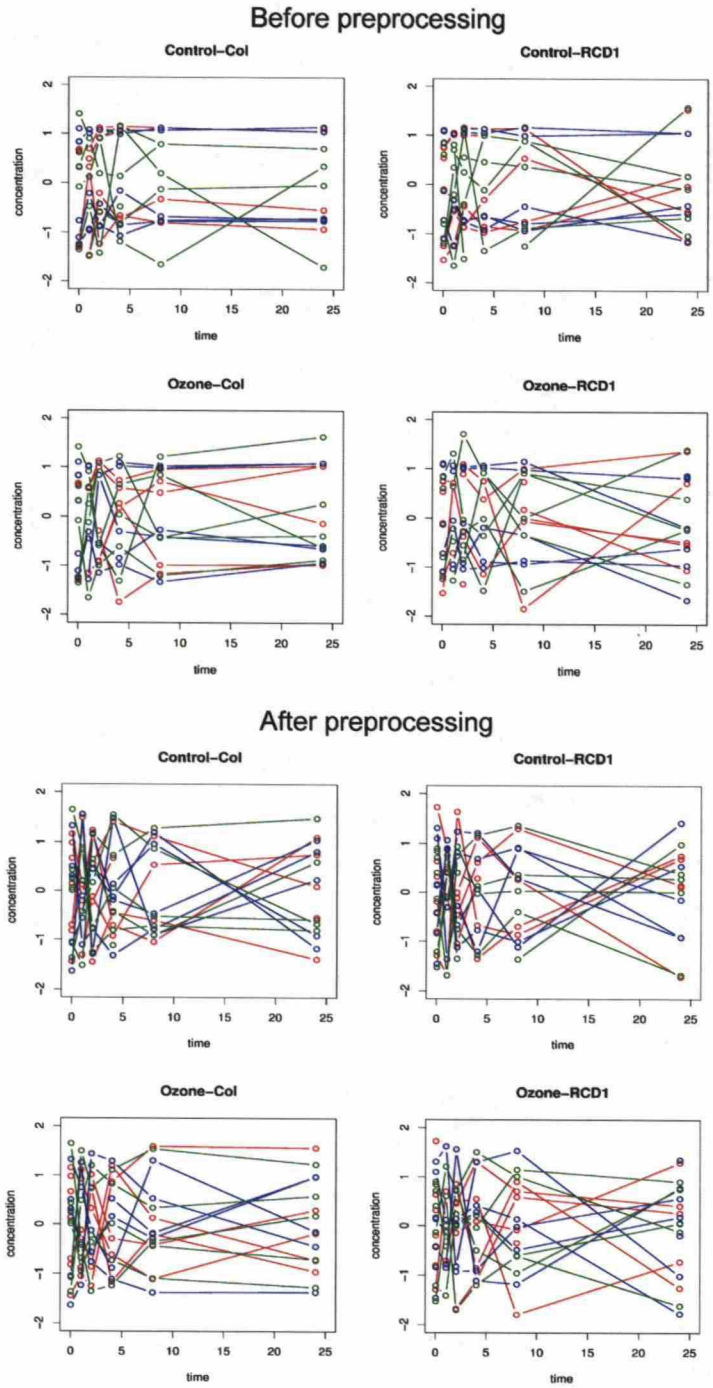


Figure A1: The effect of pre-processing to asparagine concentration levels.

Appendix B

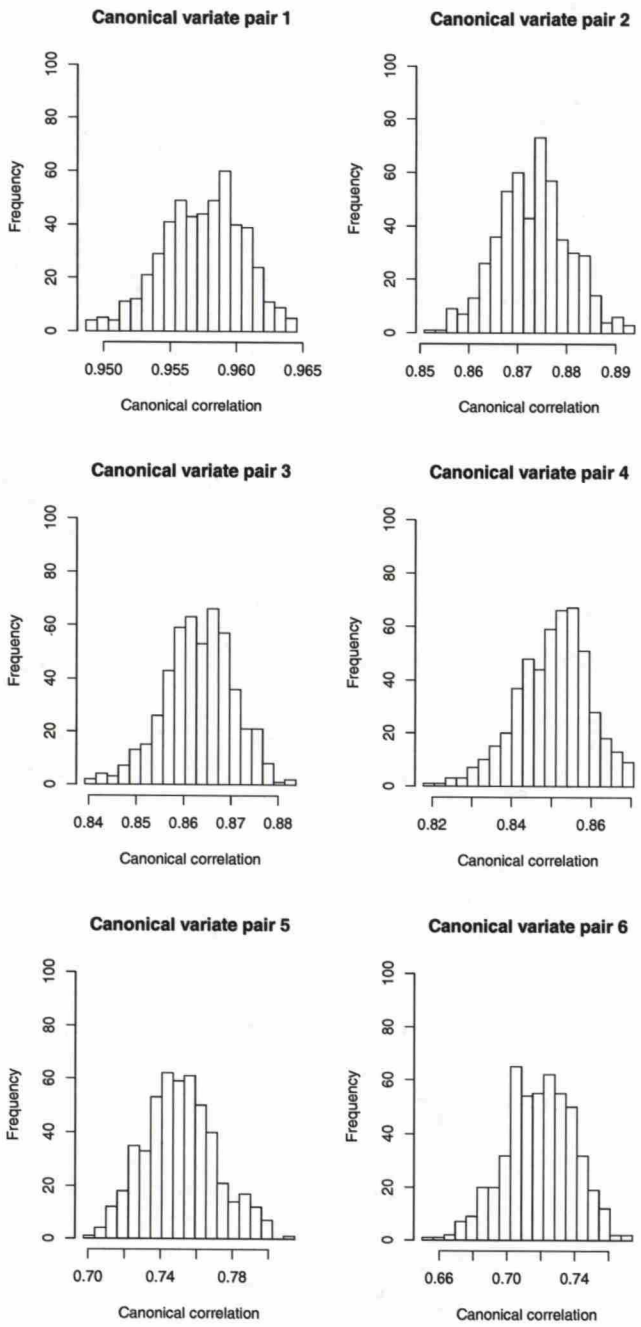


Figure B1: Posterior distributions of the canonical correlations obtained by gsCCA for canonical variate pairs 1–6.

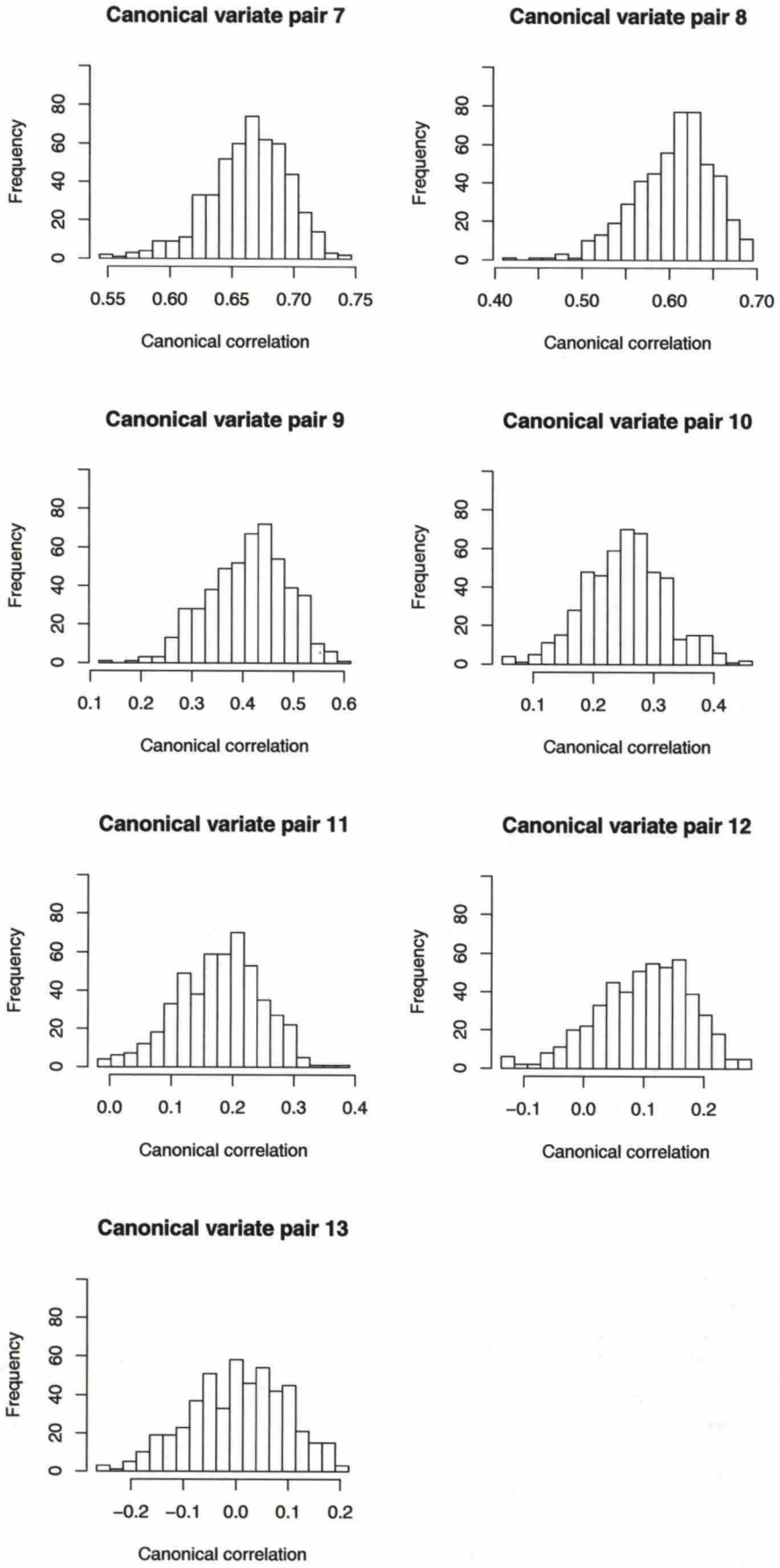


Figure B2: Posterior distributions of the canonical correlations obtained by gsCCA for canonical variate pairs 7–13.

Appendix C

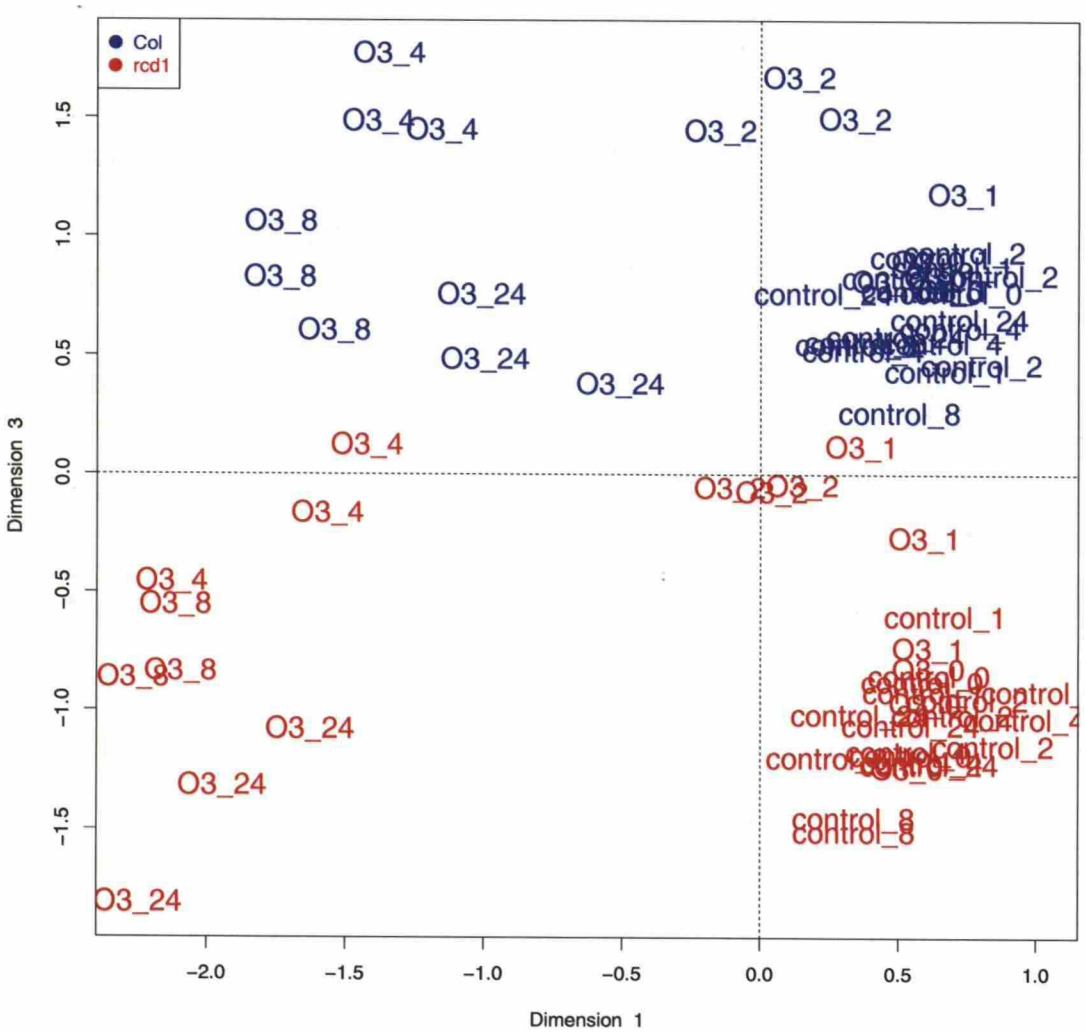


Figure C1: Samples projected on the first and third canonical variates obtained by the rrCCA method.

Appendix D

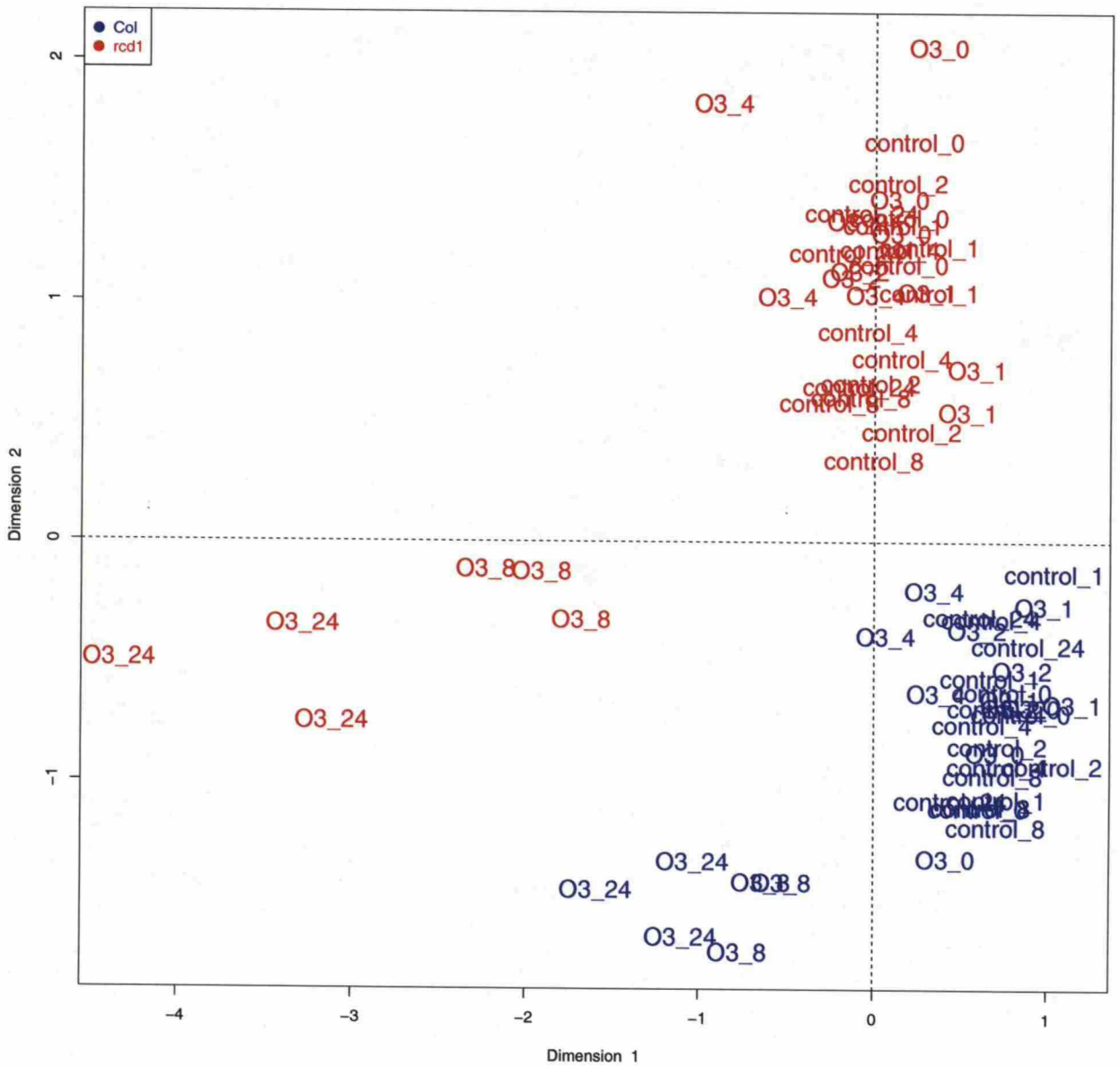


Figure D1: Samples projected on the first two canonical variates obtained by the gsCCA method.

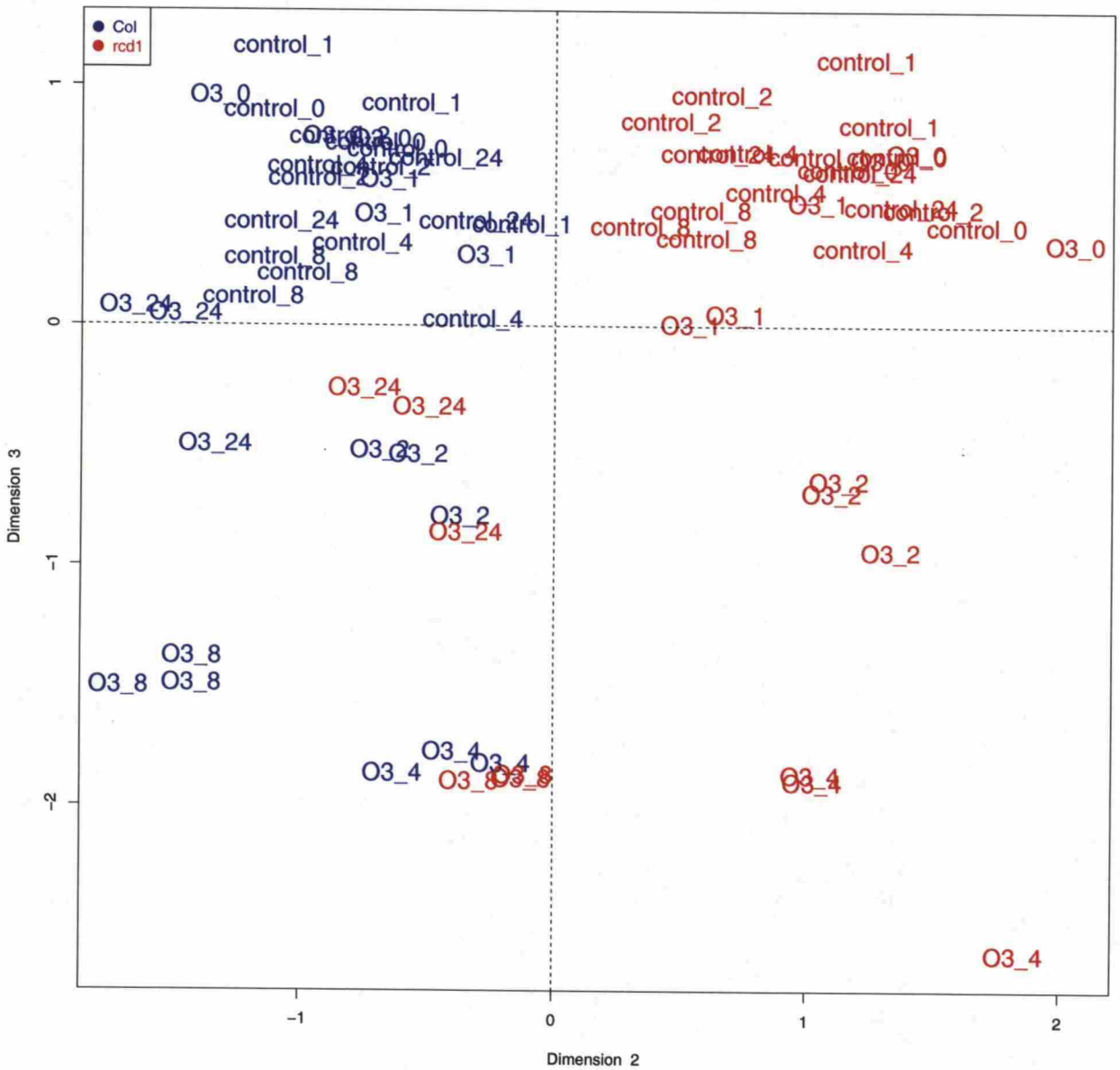


Figure D2: Samples projected on the second and third canonical variates obtained by the gsCCA method.

Appendix E

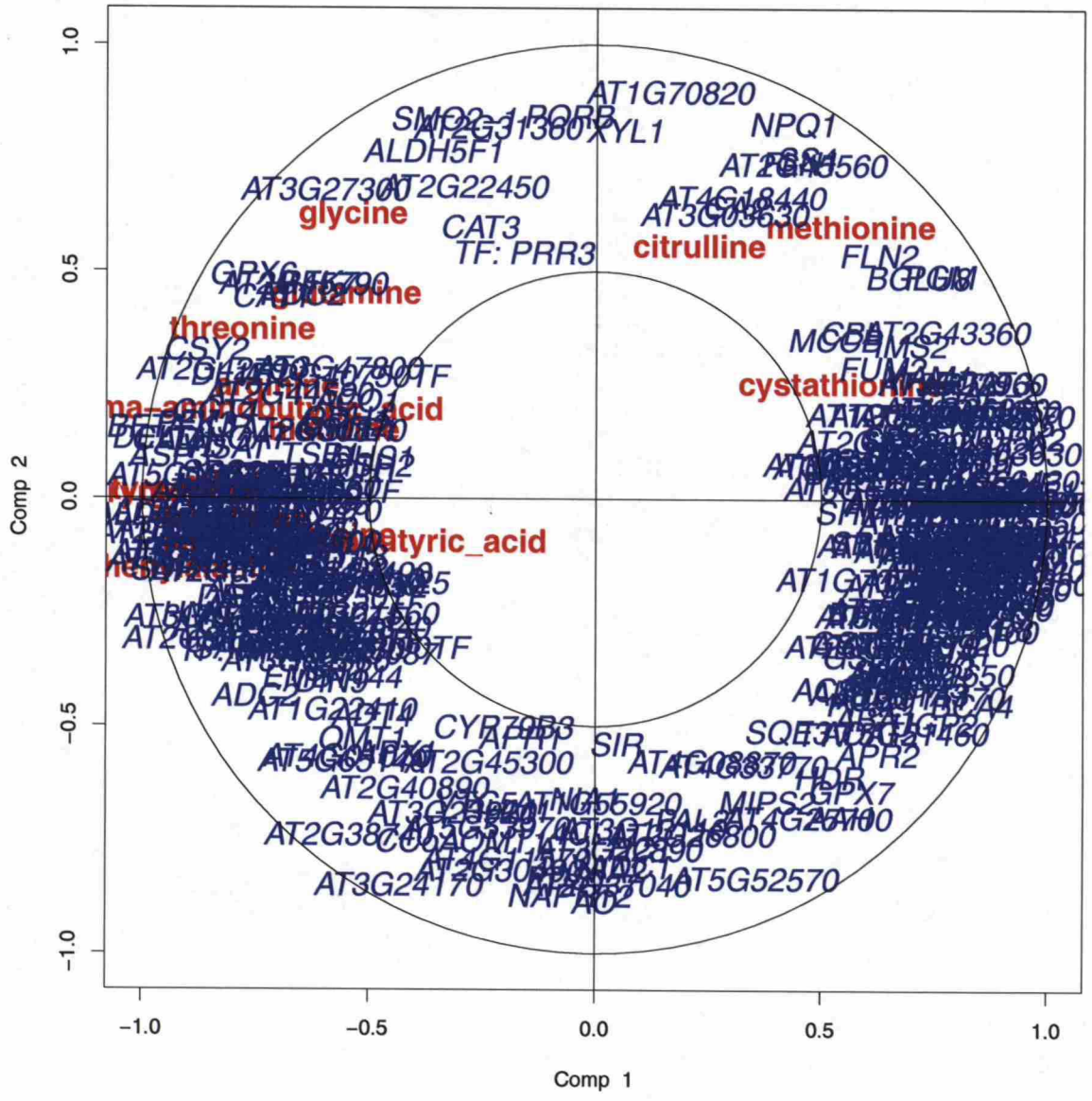


Figure E1: Canonical correlation circle when the variables are correlated on the first two canonical variates obtained by rrCCA. Metabolites are coloured red and genes are coloured blue.

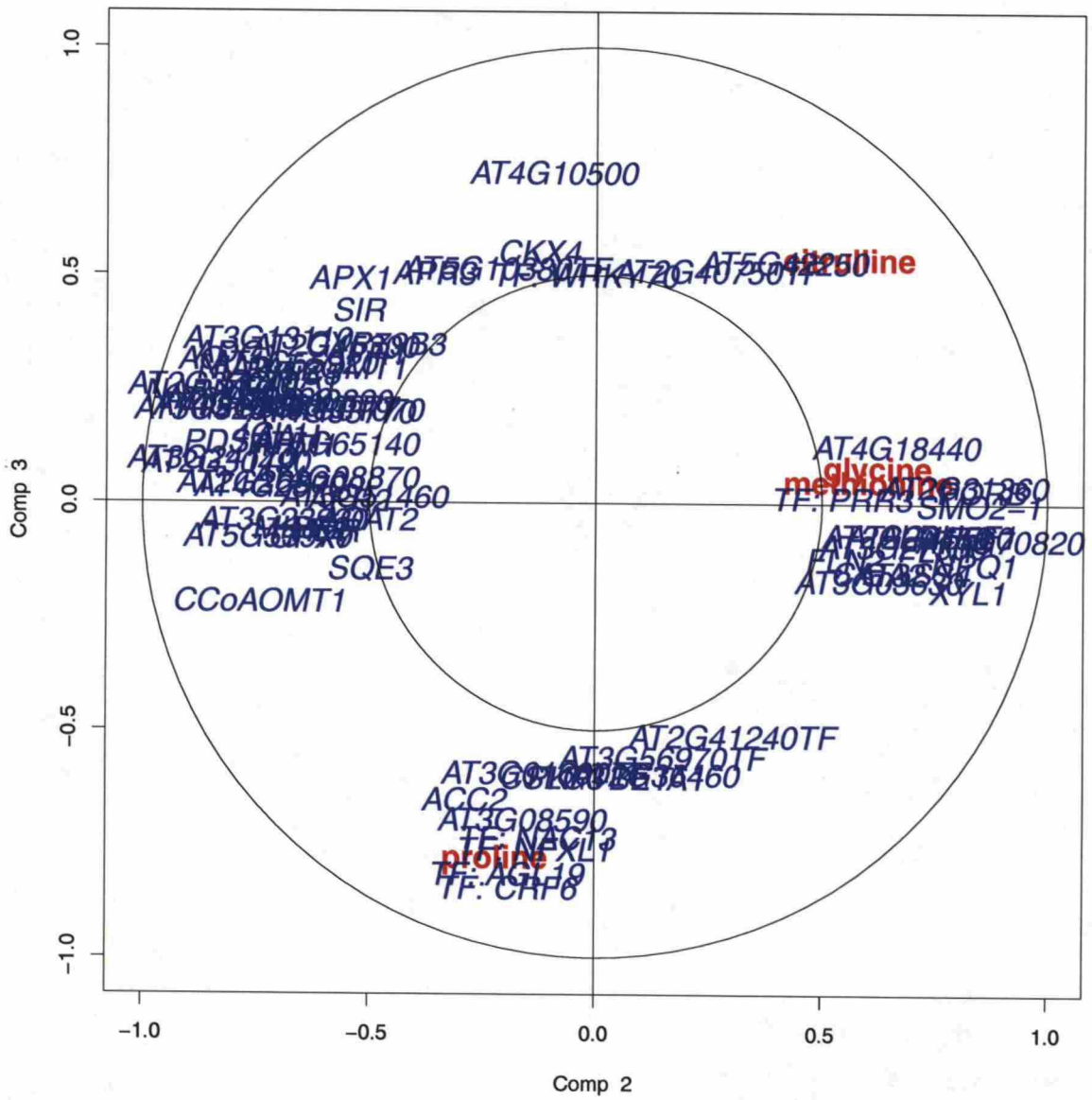


Figure E2: Canonical correlation circle when the variables are correlated on the second and third canonical variates obtained by rrCCA. Metabolites are coloured red and genes are coloured blue.

Appendix F

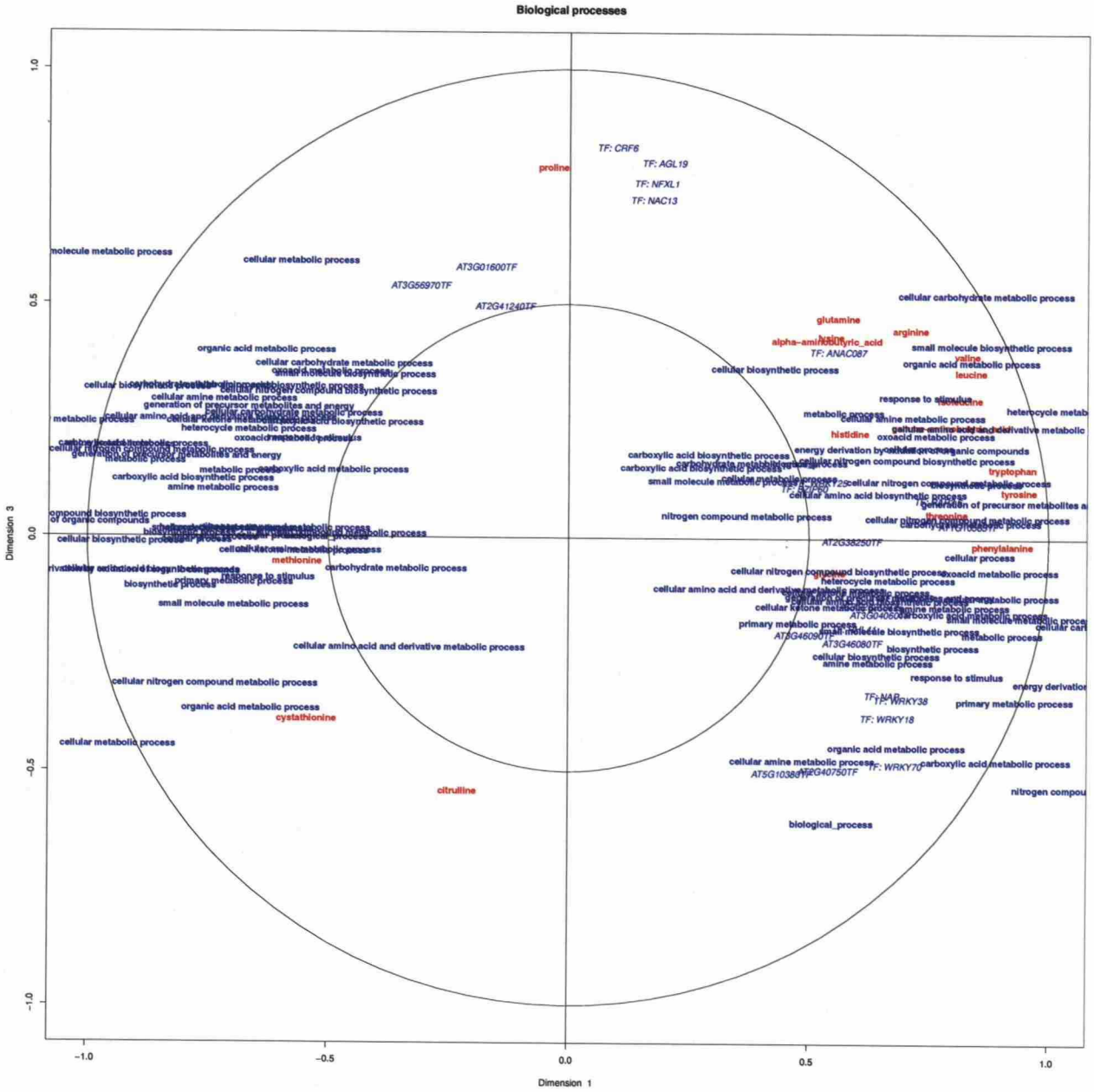


Figure F1: Enriched biological process GO classes on the first canonical component obtained by the rrCCA method. Metabolites are coloured red and genes and enriched gene sets are coloured blue.

Appendix G

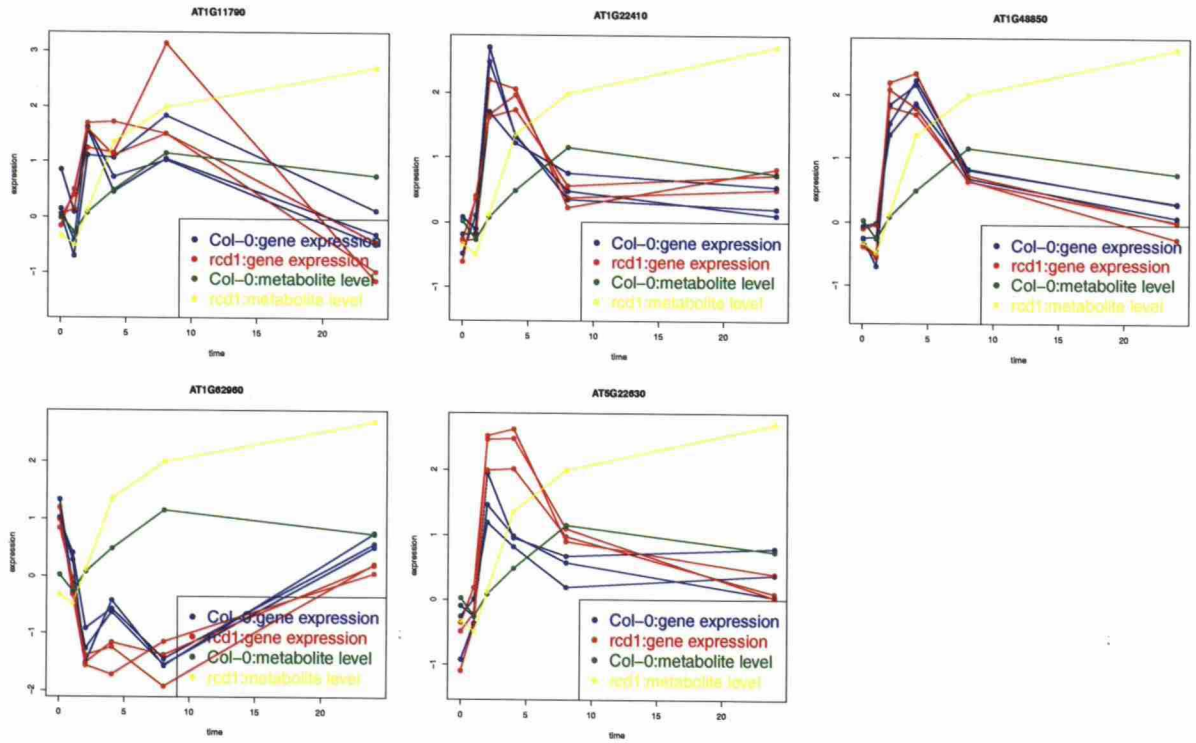


Figure G1: The concentration of metabolite phenylalanine, and expression of genes belonging to biological process GO class cellular amino acid and derivative metabolic process, and involved in phenylalanine biosynthesis. Enzymes corresponding to genes: AT1G11790, arogenate dehydratase; AT1G22410, 3-deoxy-7-phosphoheptulonate synthase; AT1G48850, chorismate synthase; AT1G62960, aspartate transaminase; AT5G22630, arogenate dehydratase. Arogenate dehydratases are involved in phenylalanine biosynthesis, aspartate transaminase is involved in phenylalanine degradation, and 3-deoxy-7-phosphoheptulonate synthase and chorismate synthase are involved in chorismate biosynthesis; chorismate is a precursor of phenylalanine, tryptophan, tyrosine and salicylic acid.

Appendix H

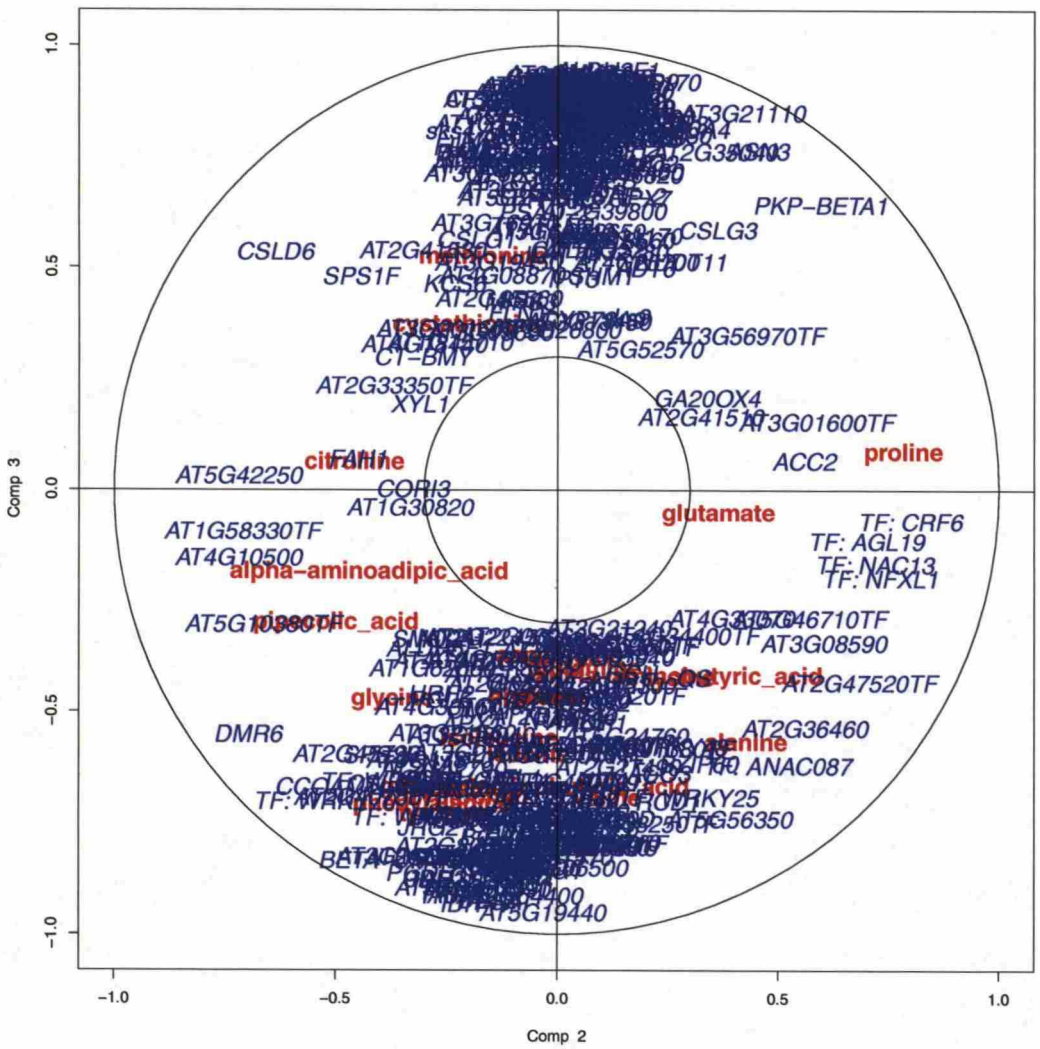


Figure H1: Canonical correlation circle when the variables are correlated on the second and third canonical variates obtained using gsCCA. Metabolites are coloured red and genes are coloured blue.

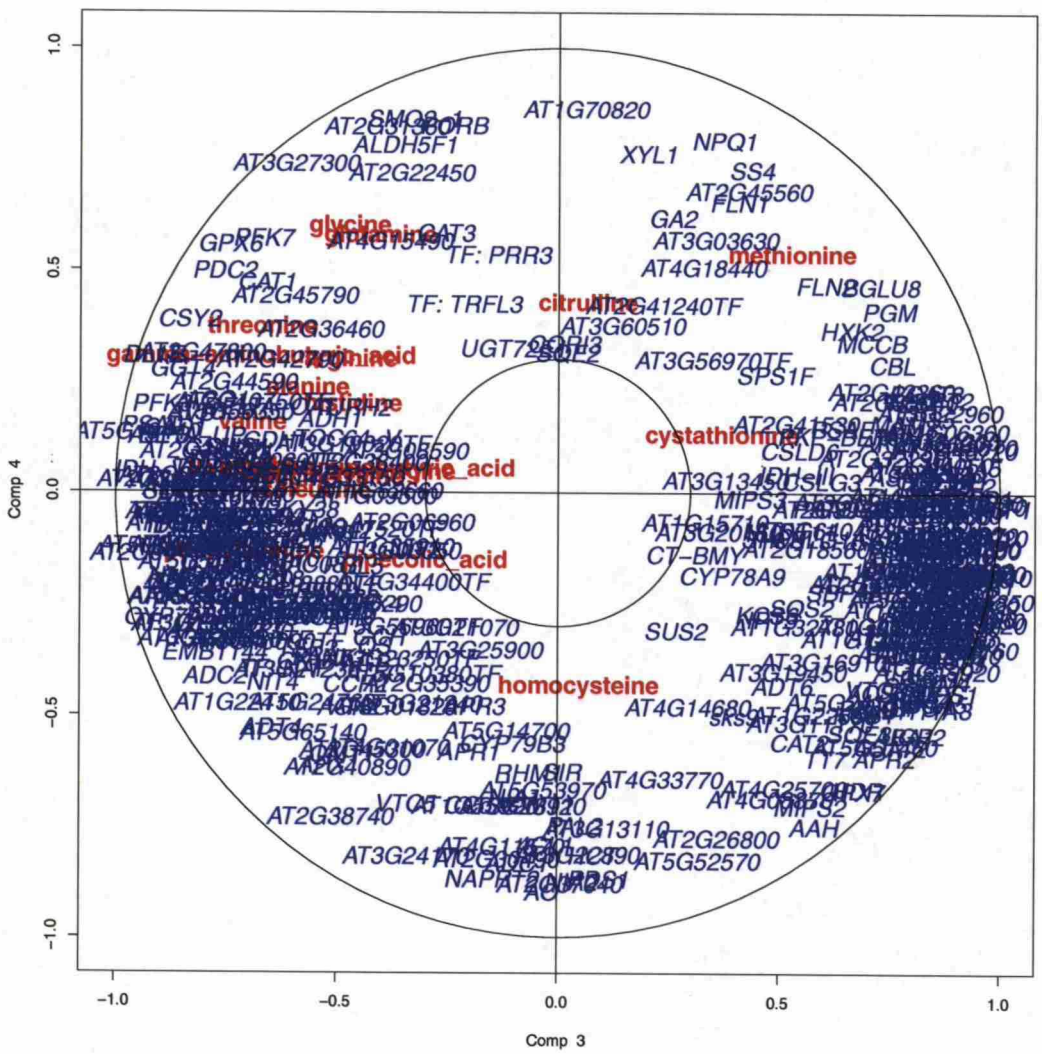


Figure H2: Canonical correlation circle when the variables are correlated on the third and fourth canonical variates obtained using gsCCA. Metabolites are coloured red and genes are coloured blue.

Appendix I

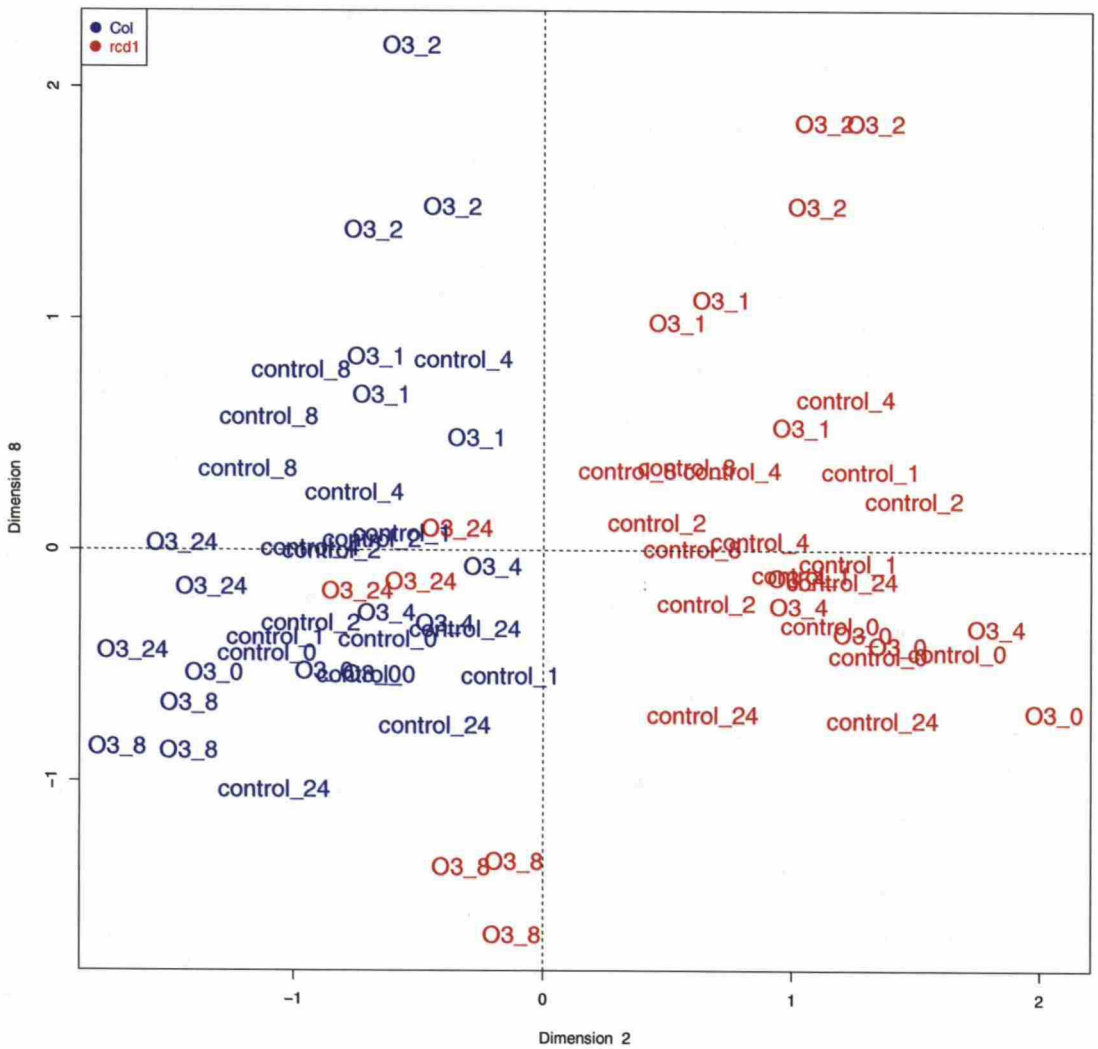


Figure I1: Samples projected on the second and eigth canonical variates obtained by gsCCA.

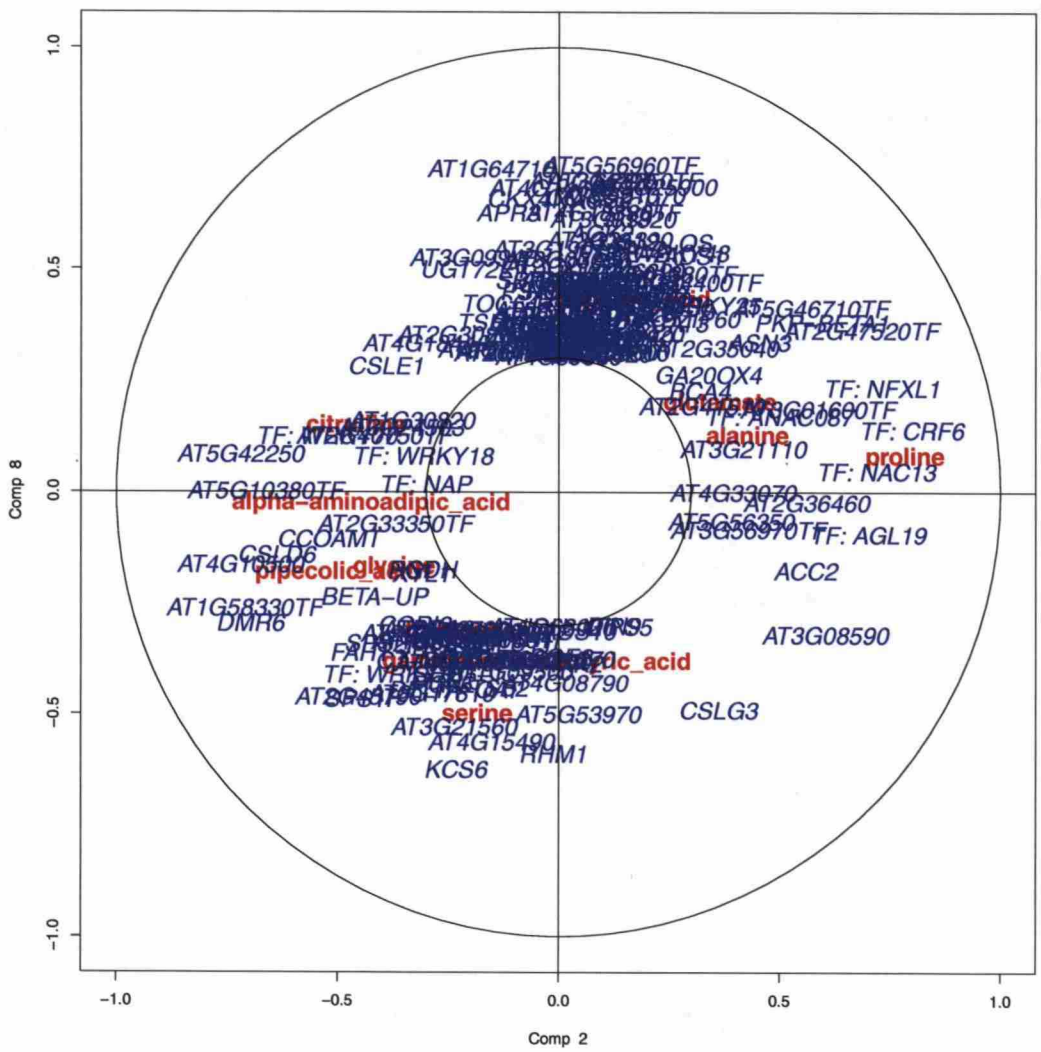


Figure I2: Canonical correlation circle when the variables are correlated on the second and 8*th* canonical variates obtained by gsCCA. Metabolites are coloured red and genes are coloured blue.

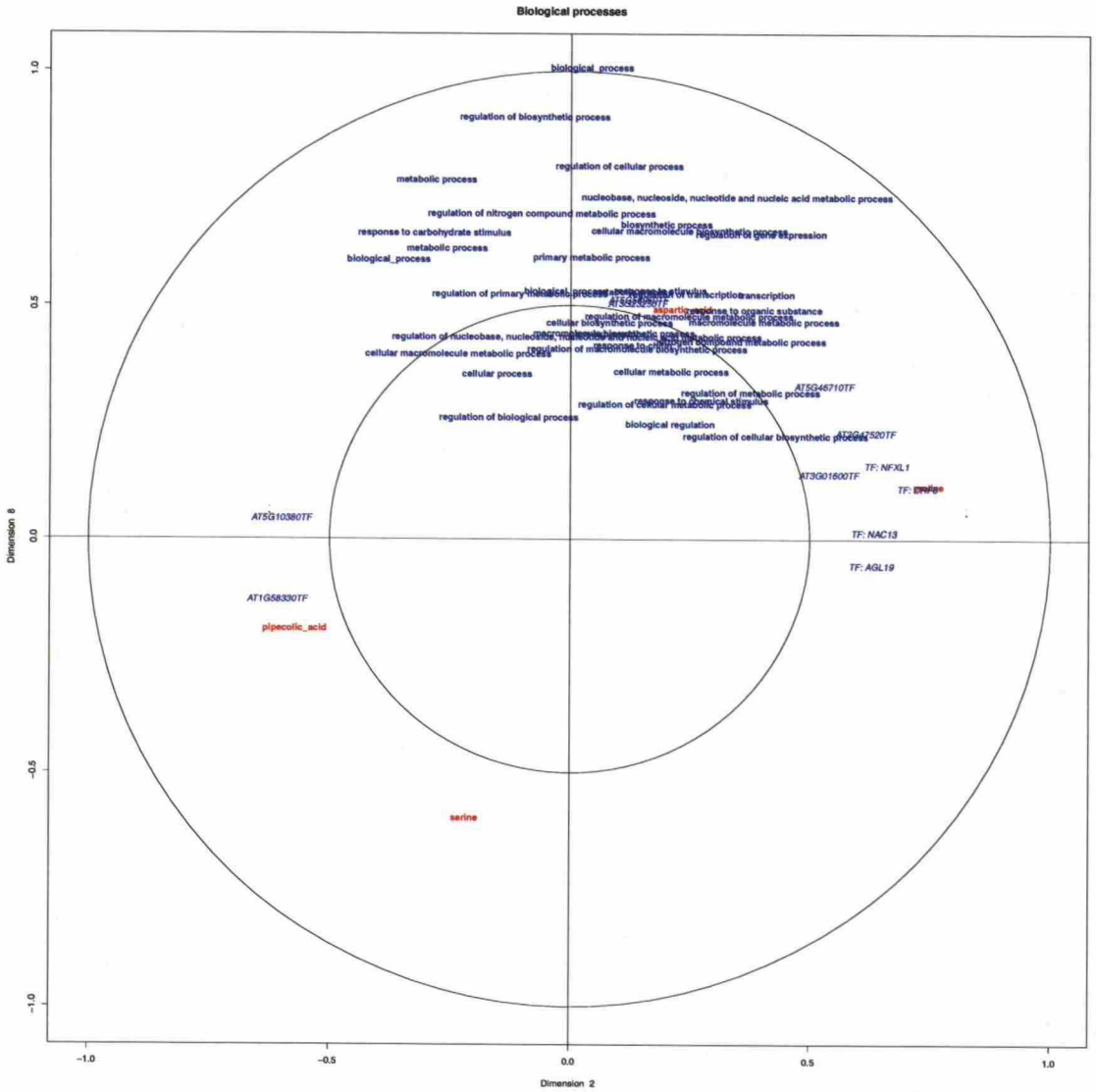


Figure I3: Canonical correlation circle when the variables are correlated on the second and $8th$ canonical variates obtained by gsCCA. Metabolites are coloured red and genes and enriched gene sets are coloured blue.

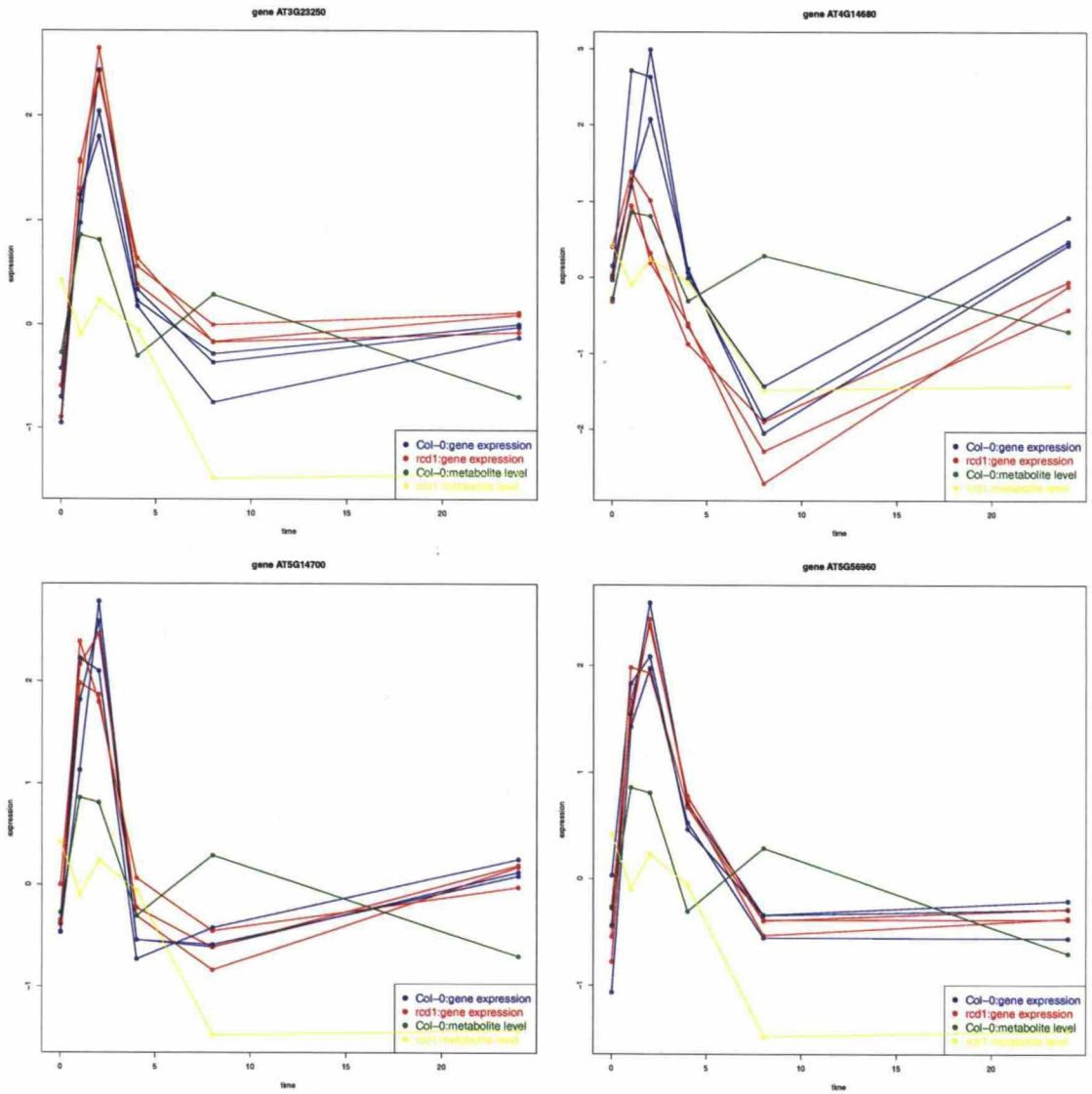


Figure I4: Expression of genes belonging to a biological process GO class cellular process, and the concentration levels of metabolite aspartic acid under ozone exposure. AT3G23250, MYB15 transcription factor; AT4G14680, ATP sulfurylase ; AT5G14700, NAD(P)-binding Rossmann-fold superfamily protein; AT5G56960, basic helix-loop-helix DNA-binding family protein.

Appendix J

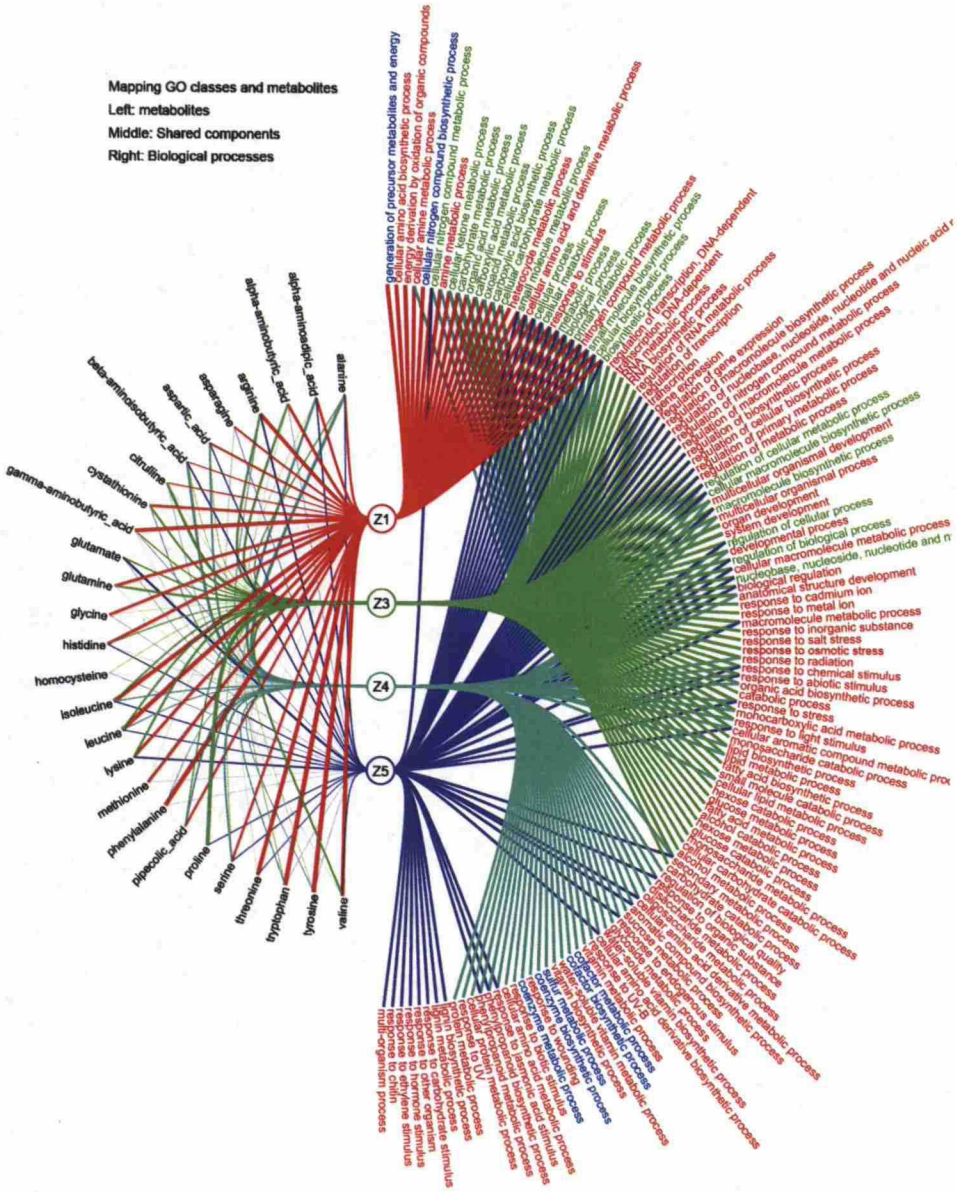


Figure J1: Enriched biological processes GO classes associated with canonical components 1, 3, 4 and 5 obtained by rrCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.

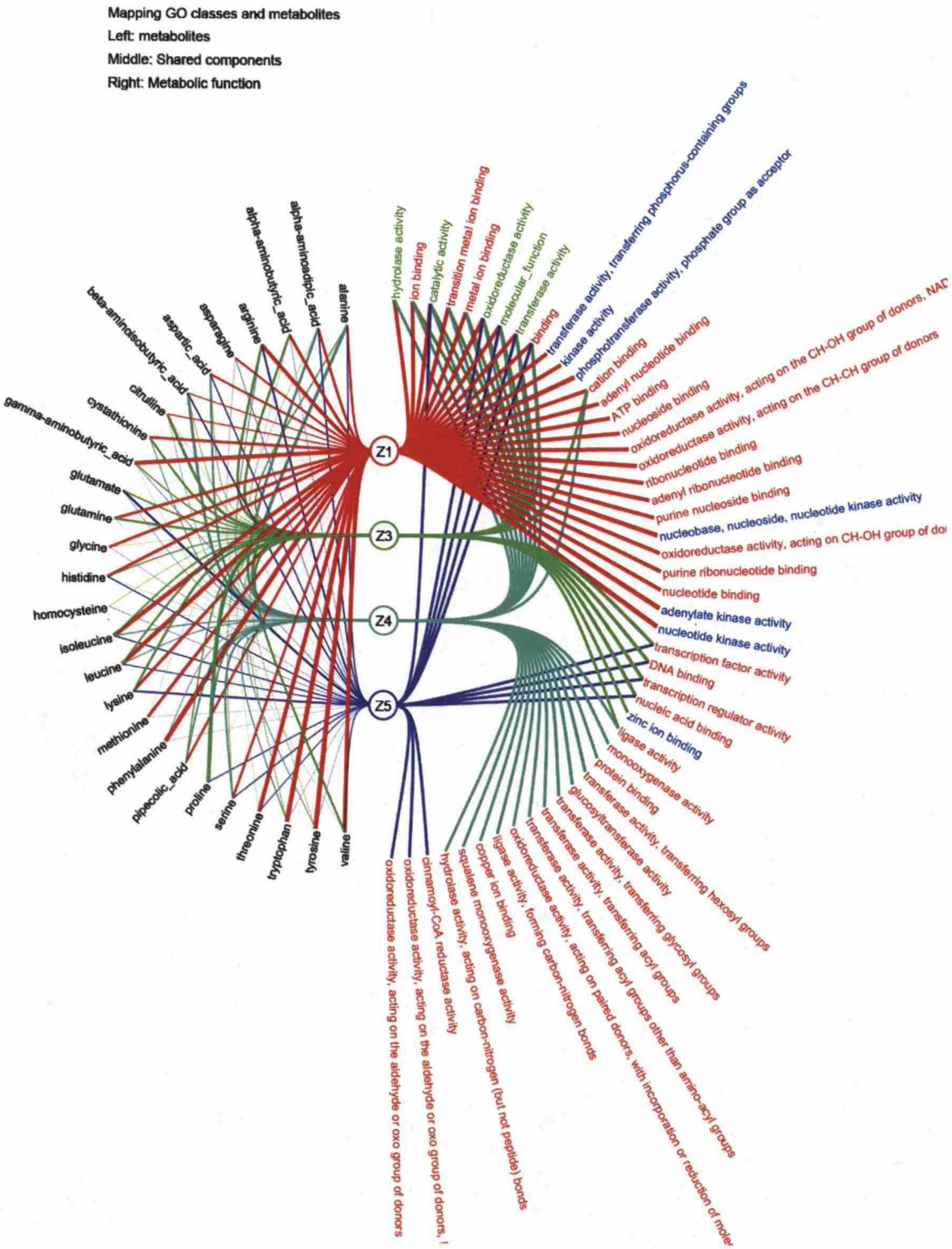


Figure J2: Enriched metabolic functions GO classes associated with canonical components 1, 3, 4 and 5 obtained by rrCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.

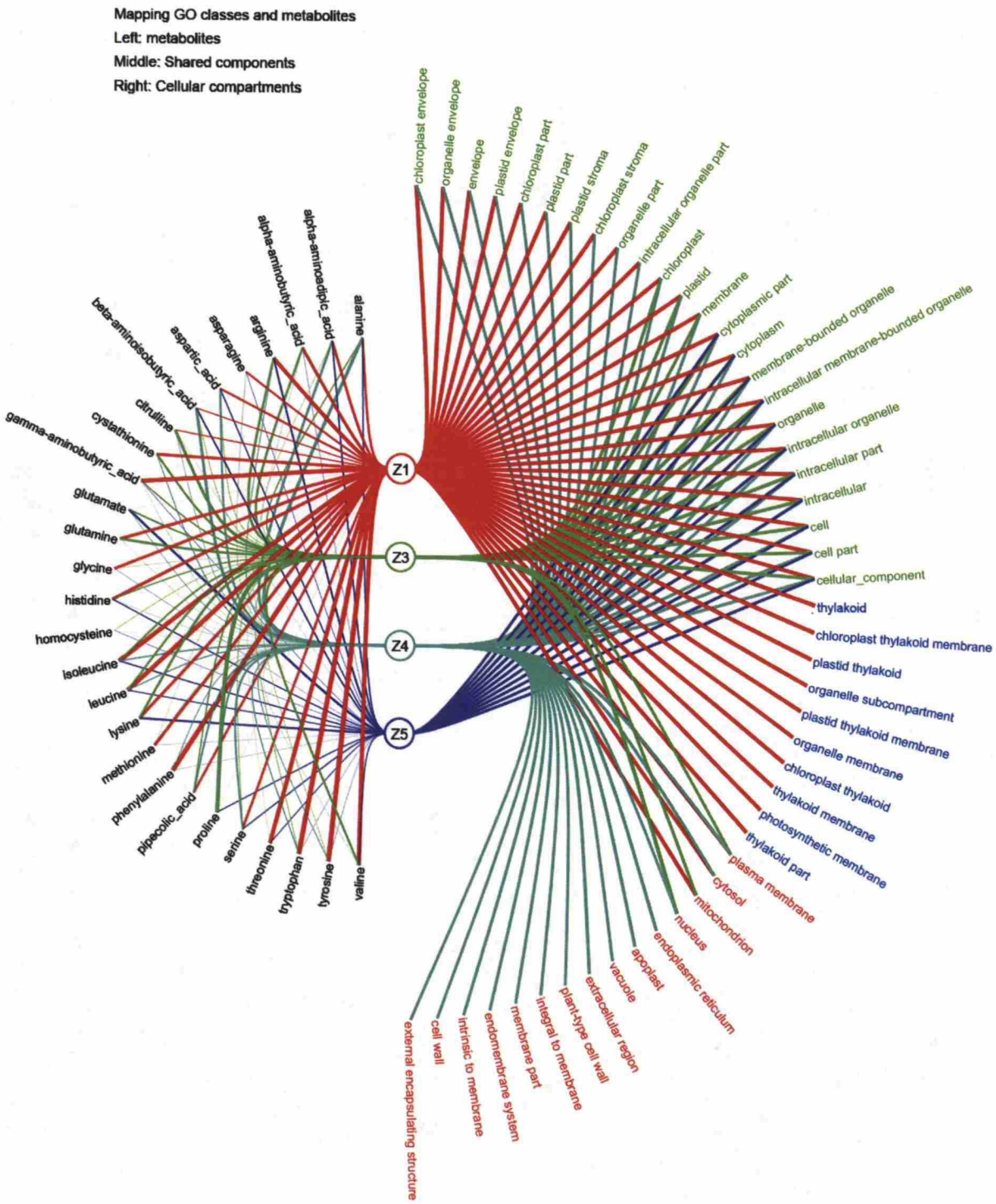


Figure J3: Enriched cellular compartment GO classes associated with canonical components 1, 3, 4 and 5 obtained by rrCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.

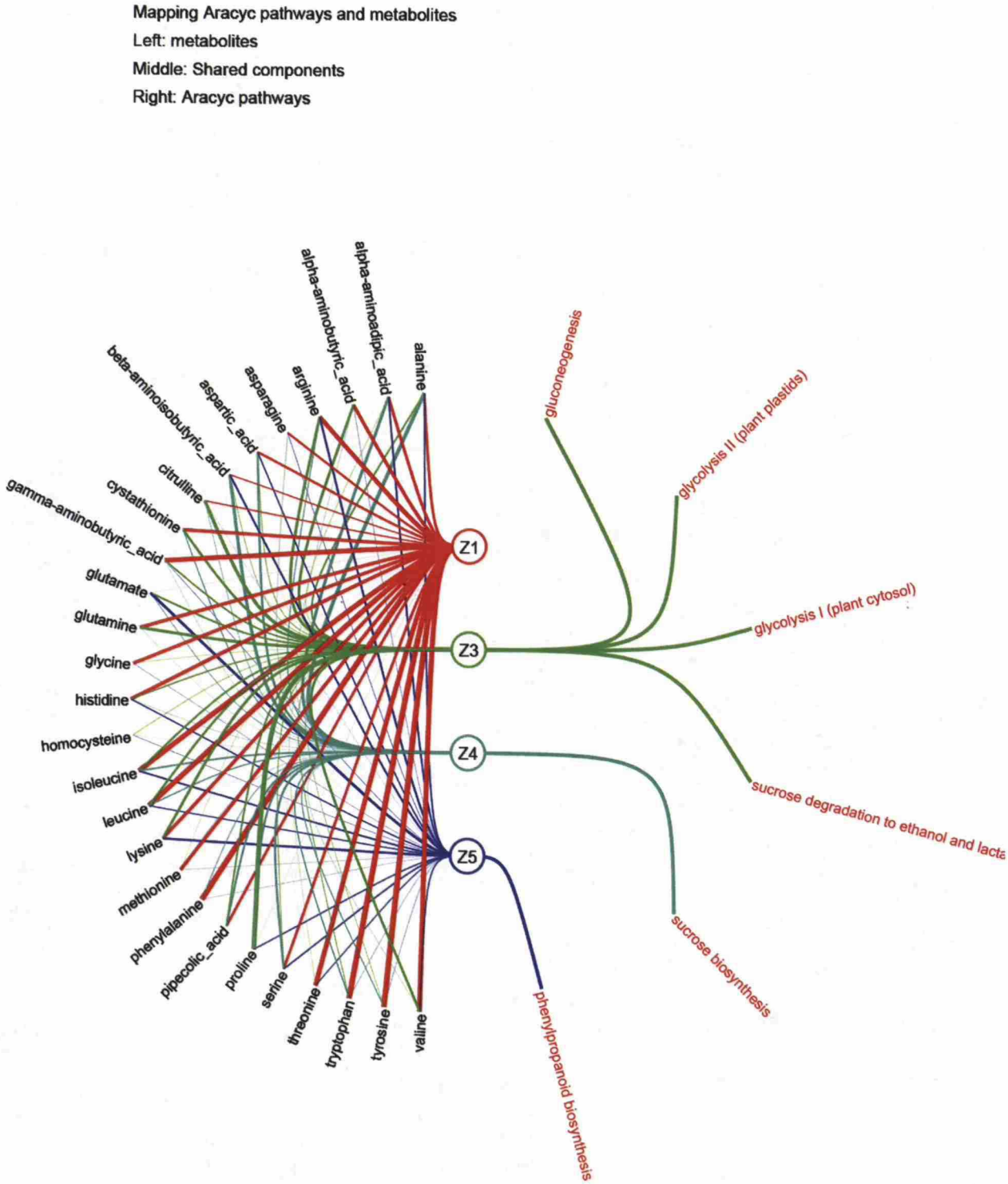


Figure J4: Enriched AraCyc pathways associated with canonical components 1, 3, 4 and 5 obtained by rrCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.

Appendix K

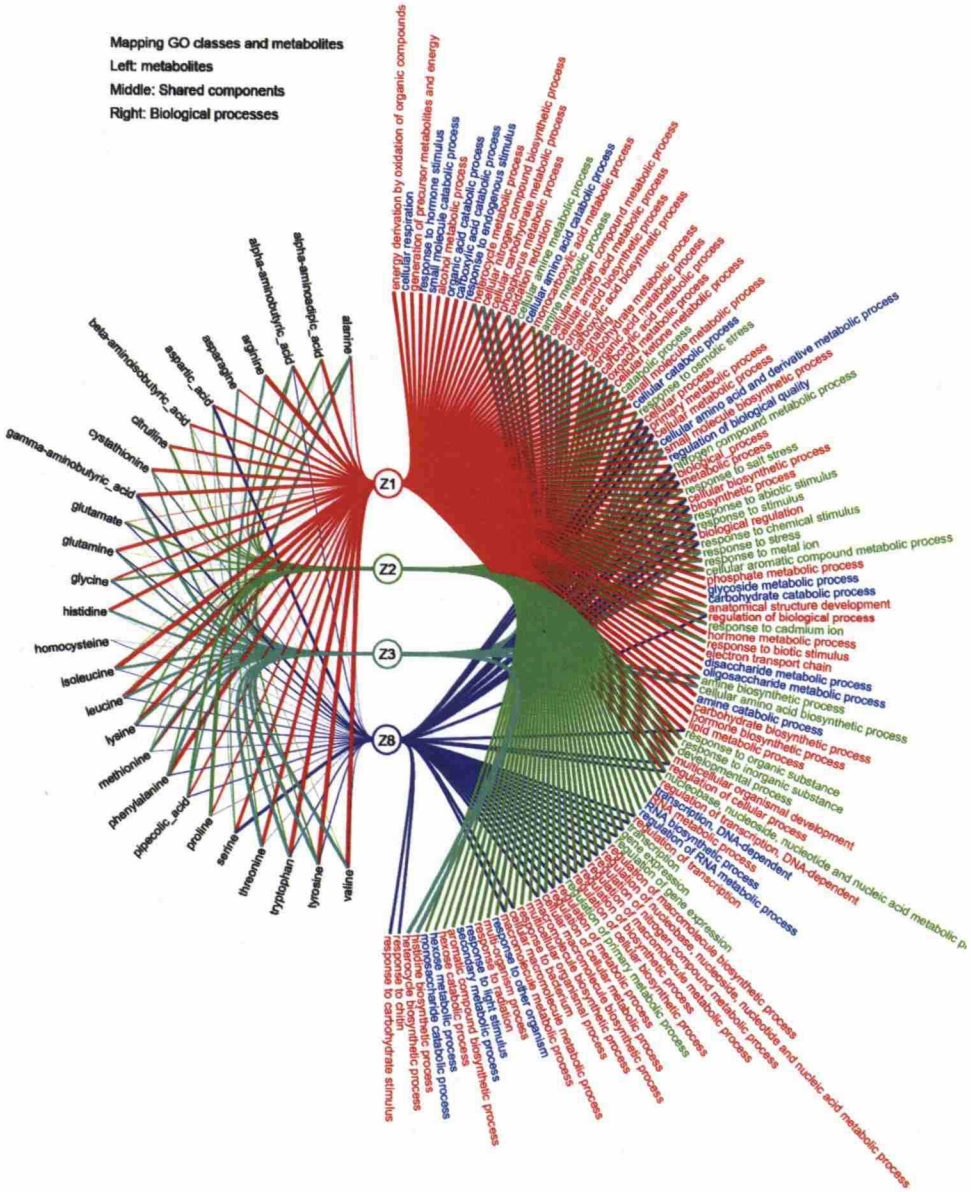


Figure K1: Enriched biological processes GO classes associated with canonical components 1, 2, 3 and 8 obtained by gsCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.

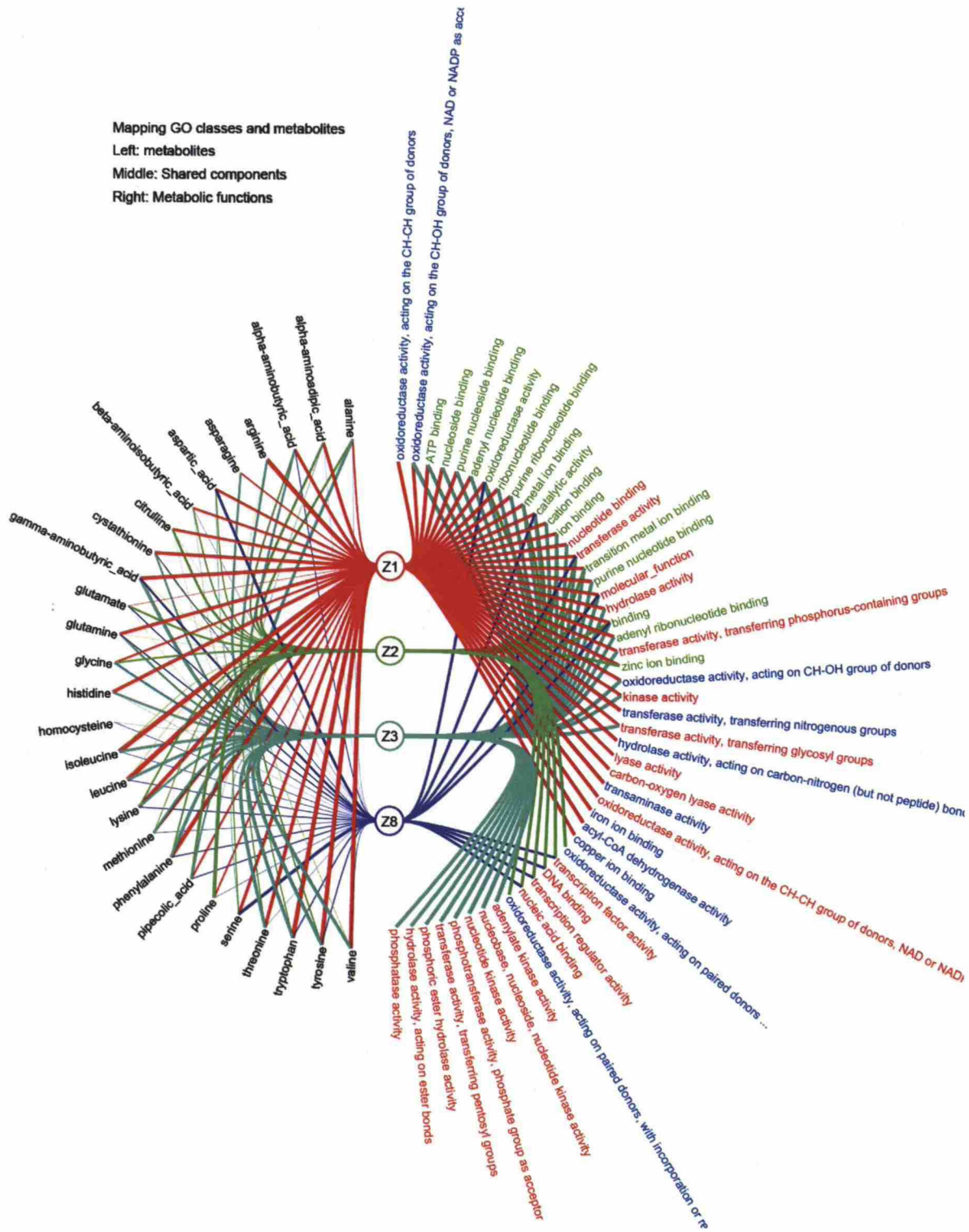


Figure K2: Enriched metabolic functions GO classes associated with canonical components 1, 2, 3, and 8 obtained by gsCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.

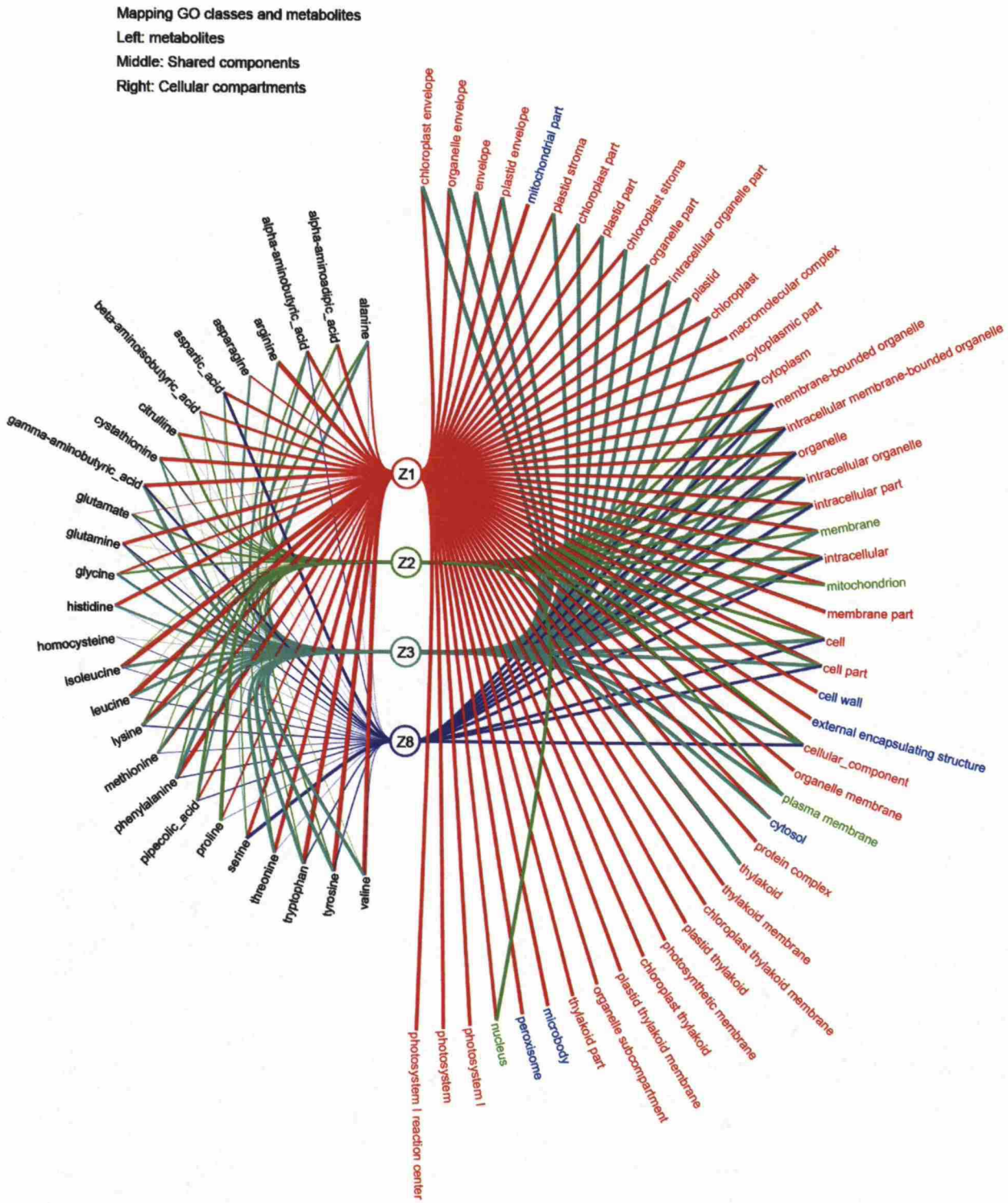


Figure K3: Enriched cellular compartment GO classes associated with canonical components 1, 2, 3, and 8 obtained by gsCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.

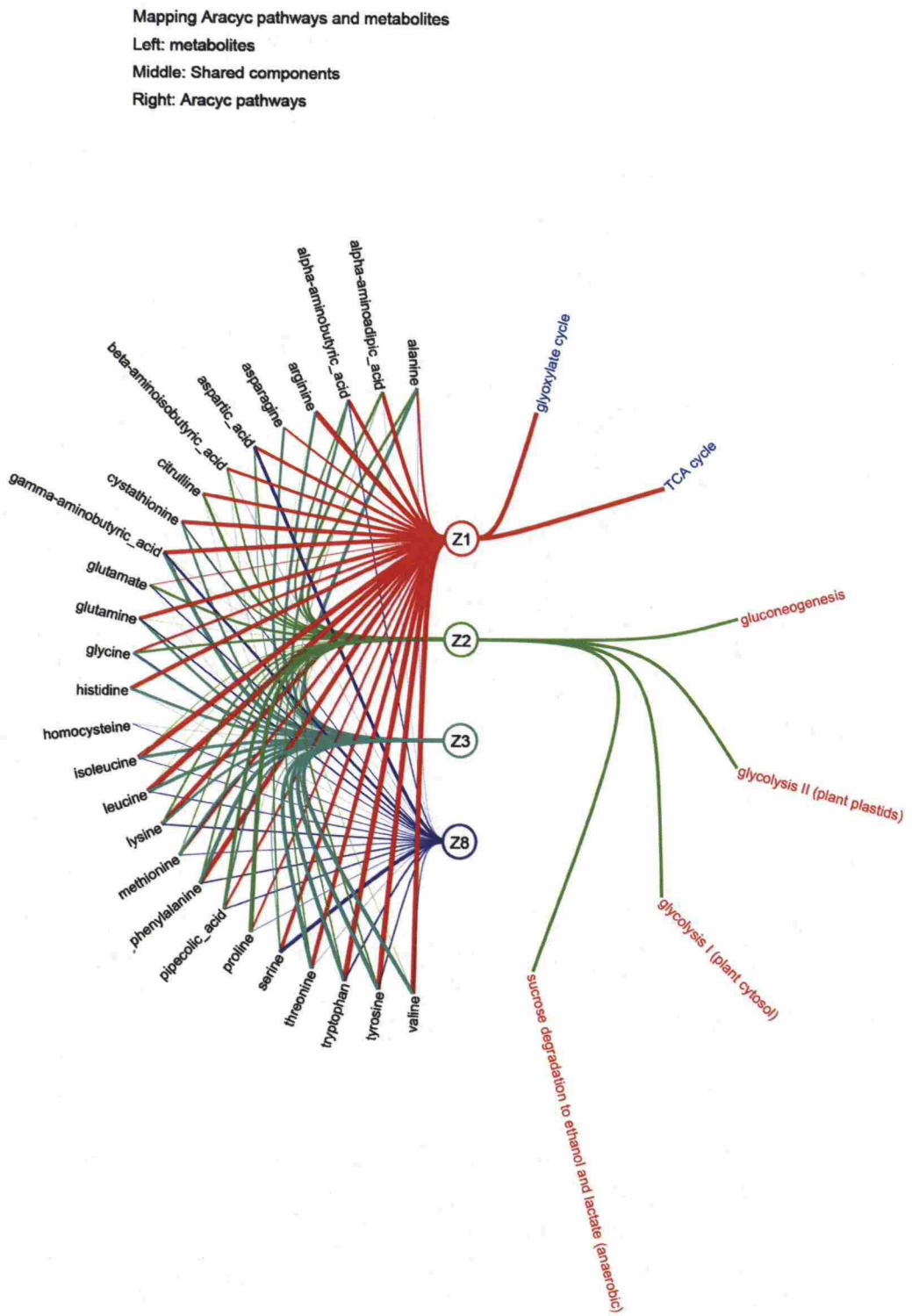


Figure K4: Enriched AraCyc pathways associated with canonical components 1, 2, 3, and 8 obtained by gsCCA. Names of the gene sets including more genes with positive canonical loading are coloured red, and names of the gene sets including more genes with negative canonical loading are coloured blue. If a gene set is associated with two or more canonical components, and the sign of the average gene set loading is different on different components, the gene set name is coloured green.